MORALS BY AGREEMENT

DAVID GAUTHIER

Professor of Philosophy
University of Pittsburgh

CLARENDON PRESS · OXFORD

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland
and associated companies in
Beirut Berlin Ibadan Nicosia

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press, New York

First published 1986
Reprinted 1987
Reprinted (new as paperback) 1987

British Library Cataloguing in Publication Data
Gauthier, David
Morals by agreement
1. Ethics
1. Title
170 BJ1012
ISBN 0–19–824746–X
ISBN 0–19–824992–6 (Pbk)

Library of Congress Cataloging in Publication Data
Gauthier, David P.
Morals by agreement
Includes index
1. Ethics. 2. Contracts. 3. Cooperation.
1. Title
BJ1012.G38 1985 171'.5 85–15519
ISBN 0–19–824746–X
ISBN 0–19–824992–6 (Pbk)

Printed in Great Britain
at the University Printing House, Oxford
by David Stanford
Printer to the University

PREFACE

THE present enquiry began on a November afternoon in Los Angeles when, fumbling for words in which to express the peculiar relationship between morality and advantage, I was shown the Prisoner's Dilemma. (The unfamiliar reader will be shown it in section 3.2 of Chapter III.) Almost nineteen years later, I reflect on the course of a voyage that is not, and cannot be, completed, but that finds a temporary harbour in this book.

The Prisoner's Dilemma posed a problem, rather than solving one. The problem concerns practical rationality, understood in maximizing terms, and it is resolved, or so I now think, in Chapter VI.

It proved to be the second of three core problems that required resolution before my enquiry could issue in this book. The first was to formulate the principle of rational co-operation, which I believe is central to morality. In my account, this principle is related to a rational agreement or bargain, and I was able to develop a game-theoretic treatment of bargaining, which has evolved into Chapter V. The second was to demonstrate the rationality of complying with this principle, which turns out to be the problem of rational behaviour in a Prisoner's Dilemma. And the third was to determine the appropriate initial position from which co-operation proceeds, which requires showing the rationality of accepting a Lockean proviso on initial acquisition. (The unfamiliar reader will meet the proviso in section 3.1 of Chapter VII.) This third problem proved the most recalcitrant; from the initial idea of a contractarian moral theory, which captured my imagination in 1966, some thirteen years elapsed before the role of the proviso became clear.

During those years I published several papers, developing what I now consider proto-versions of parts of the present theory. A reader familiar with those papers will find in them both arguments and attitudes that are contradicted or modified in the present account. I should like to think that this represents progress in my enquiry.

Perhaps changes in attitude deserve a further remark. I have had, and continue to have, somewhat mixed feelings about morals by agreement. Indeed, at one time I thought of setting out much of the

present book as a study of a set of conceptual interconnections with no claim that the whole constitutes the correct moral theory. But now I am willing to make that claim. Incorporation of the Lockean proviso, and of the idea of the liberal individual (in the final chapter), have alleviated some of my earlier worries. But perhaps most important, the conception of practical rationality that I accept at the root of my argument seems to me the only one capable of withstanding critical examination, and the moral theory that I then develop seems to me, in outline if not in every detail, the only one compatible with that conception of rationality. Yet, as Richard Rorty or Alasdair MacIntyre might remind me, perhaps I lack the vocabulary for talking perspicuously about morality.

The first draft of most of the book was written in Aix-en-Provence in 1979–80; I am grateful for financial support during that year from the University of Toronto and from the Social Sciences and Humanities Research Council of Canada. The second draft of Chapters II–IX was written in Toronto in the spring of 1982; the final chapters refused to fall into place until the summer of 1983. The present draft was written in Pittsburgh from January to May 1984; I am grateful to the University of Pittsburgh for releasing me from teaching duties during that term, and to Ruth Durst for her typing.

Were I to attempt to thank by name all those individuals who have contributed to an enquiry as extended as this one, my list would inevitably be marred by omission. Many, many colleagues and students, at the Universities of Toronto and Pittsburgh, at other universities where I have visited or read papers, and at professional meetings, have helped me to rethink my ideas and to reformulate my thoughts. To all, my gratitude. To Clark Glymour, who pointed out that I could avoid the horrendous 's/he' and 'him/her' of earlier drafts by resorting to an appropriately programmed randomizing device, I am sure that I can add the gratitude of readers to my own. And to Howard Sobel, who among those who have provided philosophical aid occupies the first place, for it was Howard who showed me the Prisoner's Dilemma and introduced me to the basic ideas of the theory of games, mere gratitude would not be enough.

And so I come to an end, aware that it is also a beginning, for I shall surely find myself embarked again on the quest to understand how morality and rationality are related.

DAVID GAUTHIER

Pittsburgh
24 May 1984

CONTENTS

I

OVERVIEW OF A THEORY

1 What theory of morals can ever serve any useful purpose, unless it can show that all the duties it recommends are also the true interest of each individual?[1] David Hume, who asked this question, seems mistaken; such a theory would be too useful. Were duty no more than interest, morals would be superfluous. Why appeal to right or wrong, to good or evil, to obligation or to duty, if instead we may appeal to desire or aversion, to benefit or cost, to interest or to advantage? An appeal to morals takes its point from the failure of these latter considerations as sufficient guides to what we ought to do. The unphilosophical poet Ogden Nash grasped the assumptions underlying our moral language more clearly than the philosopher Hume when he wrote:

> 'O Duty!
> Why hast thou not the visage of a sweetie or a cutie?'[2]

We may lament duty's stern visage but we may not deny it. For it is only as we believe that some appeals do, alas, override interest or advantage that morality becomes our concern.

But if the language of morals is not that of interest, it is surely that of reason. What theory of morals, we might better ask, can ever serve any useful purpose, unless it can show that all the duties it recommends are also truly endorsed in each individual's reason? If moral appeals are entitled to some practical effect, some influence on our behaviour, it is not because they whisper invitingly to our desires, but because they convince our intellect. Suppose we should find, as Hume himself believes, that reason is impotent in the sphere of action apart from its role in deciding matters of fact.[3] Or suppose we should find that reason is no more than the handmaiden of

[1] See David Hume, *An Enquiry concerning the Principles of Morals*, sect. ix, pt. ii, in L. A. Selby-Bigge (ed.), *Enquiries concerning Human Understanding and concerning the Principles of Morals*, 3rd edn. (Oxford, 1975), p. 280.

[2] Ogden Nash, 'Kind of an Ode to Duty', *I Wouldn't Have Missed It: Selected Poems of Ogden Nash* (Boston, 1975), p. 141.

[3] See David Hume, *A Treatise of Human Nature*, bk. ii, pt. iii, sect. iii, ed. L. A. Selby-Bigge (Oxford, 1888), pp. 413–18.

interest, so that in overriding advantage a moral appeal must also contradict reason. In either case we should conclude that the moral enterprise, as traditionally conceived, is impossible.

To say that our moral language assumes a connection with reason is not to argue for the rationality of our moral views, or of any alternative to them. Moral language may rest on a false assumption.[4] If moral duties are rationally grounded, then the emotivists, who suppose that moral appeals are no more than persuasive, and the egoists, who suppose that rational appeals are limited by self-interest, are mistaken.[5] But are moral duties rationally grounded? This we shall seek to prove, showing that reason has a practical role related to but transcending individual interest, so that principles of action that prescribe duties overriding advantage may be rationally justified. We shall defend the traditional conception of morality as a rational constraint on the pursuit of individual interest.

Yet Hume's mistake in insisting that moral duties must be the true interest of each individual conceals a fundamental insight. Practical reason is linked to interest, or, as we shall come to say, to individual utility, and rational constraints on the pursuit of interest have themselves a foundation in the interest they constrain. Duty overrides advantage, but the acceptance of duty is truly advantageous. We shall find this seeming paradox embedded in the very structure of interaction. As we come to understand this structure, we shall recognize the need for restraining each person's pursuit of her own utility, and we shall examine its implications for both our principles of action and our conception of practical rationality. Our enquiry will lead us to the rational basis for a morality, not of absolute standards, but of agreed constraints.

2.1 We shall develop a theory of morals. Our concern is to provide a justificatory framework for moral behaviour and principles, not an explanatory framework. Thus we shall develop a normative theory. A complete philosophy of morals would need to explain, and perhaps to defend, the idea of a normative theory. We shall not do this. But we shall exemplify normative theory by sketching the theory of rational choice. Indeed, we shall do more. We shall develop a theory of morals as part of the theory of rational

[4] Thus one might propose an error theory of moral language; for the idea of an error theory, see J. L. Mackie, *Ethics: Inventing Right and Wrong* (Harmondsworth, Middx., 1977), ch. 1, esp. pp. 35, 48–9.

[5] The idea that moral appeals are persuasive is developed by C. L. Stevenson; see *Ethics and Language* (New Haven, 1944), esp. chs. vi, ix.

choice. We shall argue that the rational principles for making choices, or decisions among possible actions, include some that constrain the actor pursuing his own interest in an impartial way. These we identify as moral principles.

The study of choice begins from the stipulation of clear conceptions of value and rationality in a form applicable to choice situations.[6] The theory then analyses the structure of these situations so that, for each type of structure distinguished, the conception of rationality may be elaborated into a set of determinate conditions on the choice among possible actions. These conditions are then expressed as precise principles of rational behaviour, serving both for prescription and for critical assessment. Derivatively, the principles also have an explanatory role in so far as persons actually act rationally.

The simplest, most familiar, and historically primary part of this study constitutes the core of classical and neo-classical economic theory, which examines rational behaviour in those situations in which the actor knows with certainty the outcome of each of his possible actions. The economist does of course offer to explain behaviour, and much of the interest of her theory depends on its having explanatory applications, but her explanations use a model of ideal interaction which includes the rationality of the actors among its assumptions. Thus economic explanation is set within a normative context. And the role of economics in formulating and evaluating policy alternatives should leave us in no doubt about the deeply prescriptive and critical character of the science.

The economist formulates a simple, maximizing conception of practical rationality, which we shall examine in Chapter II. But the assumption that the outcome of each possible choice can be known with certainty seriously limits the scope of economic analysis and the applicability of its account of reason. Bayesian decision theory relaxes this assumption, examining situations with choices involving risk or uncertainty. The decision theorist is led to extend the economist's account of reason, while preserving its fundamental identification of rationality with maximization.

Both economics and decision theory are limited in their analysis of

[6] Our sketch of rational choice owes much to J. C. Harsanyi; see 'Advances in Understanding Rational Behavior', in *Essays on Ethics, Social Behavior, and Scientific Explanation* (Dordrecht, 1976), pp. 89–98, and 'Morality and the theory of rational behaviour', in A. Sen and B. Williams (eds.), *Utilitarianism and beyond* (Cambridge, 1982), pp. 42–4.

interaction, since both consider outcomes only in relation to the choices of a single actor, treating the choices of others as aspects of that actor's circumstances. The theory of games overcomes this limitation, analysing outcomes in relation to sets of choices, one for each of the persons involved in bringing about the outcome. It considers the choices of an actor who decides on the basis of expectations about the choices of others, themselves deciding on the basis of expectations about his choice. Since situations involving a single actor may be treated as limiting cases of interaction, game theory aims at an account of rational behaviour in its full generality. Unsurprisingly, achievements are related inversely to aims; as a study of rational behaviour under certainty economic theory is essentially complete, whereas game theory is still being developed. The theory of rational choice is an ongoing enterprise, extending a basic understanding of value and rationality to the formulation of principles of rational behaviour in an ever wider range of situations.

2.2 Rational choice provides an exemplar of normative theory. One might suppose that moral theory and choice theory are related only in possessing similar structures. But as we have said, we shall develop moral theory as part of choice theory. Those acquainted with recent work in moral philosophy may find this a familiar enterprise; John Rawls has insisted that the theory of justice is 'perhaps the most significant part, of the theory of rational choice', and John Harsanyi explicitly treats ethics as part of the theory of rational behaviour.[7] But these claims are stronger than their results warrant. Neither Rawls nor Harsanyi develops the deep connection between morals and rational choice that we shall defend. A brief comparison will bring our enterprise into sharper focus.

Our claim is that in certain situations involving interaction with others, an individual chooses rationally only in so far as he constrains his pursuit of his own interest or advantage to conform to principles expressing the impartiality characteristic of morality. To choose rationally, one must choose morally. This is a strong claim. Morality, we shall argue, can be generated as a rational constraint from the non-moral premisses of rational choice. Neither Rawls nor Harsanyi makes such a claim. Neither Rawls nor Harsanyi treats moral principles as a subset of rational principles for choice.

Rawls argues that the principles of justice are the objects of a

[7] J. Rawls, *A Theory of Justice* (Cambridge, Mass., 1971), p. 16; J. Harsanyi, 'Morality and the theory of rational behaviour', p. 42.

rational choice—the choice that any person would make, were he called upon to select the basic principles of his society from behind a 'veil of ignorance' concealing any knowledge of his own identity.[8] The principles so chosen are not directly related to the making of individual choices.[9] Derivatively, acceptance of them must have implications for individual behaviour, but Rawls never claims that these include rational constraints on individual choices. They may be, in Rawls's terminology, reasonable constraints, but what is reasonable is itself a morally substantive matter beyond the bounds of rational choice.[10]

Rawls's idea, that principles of justice are the objects of a rational choice, is indeed one that we shall incorporate into our own theory, although we shall represent the choice as a bargain, or agreement, among persons who need not be unaware of their identities. But this parallel between our theory and Rawls's must not obscure the basic difference; we claim to generate morality as a set of rational principles for choice. We are committed to showing why an individual, reasoning from non-moral premisses, would accept the constraints of morality on his choices.

Harsanyi's theory may seem to differ from Rawls's only in its account of the principles that a person would choose from behind a veil of ignorance; Rawls supposes that persons would choose the well-known two principles of justice, whereas Harsanyi supposes that persons would choose principles of average rule-utilitarianism.[11] But Harsanyi's argument is in some respects closer to our own; he is concerned with principles for moral choice, and with the rational way of arriving at such principles. However, Harsanyi's principles are strictly hypothetical; they govern rational choice from an impartial standpoint or given impartial preferences, and so they are principles only for someone who wants to choose morally or impartially.[12] But Harsanyi does not claim, as we do, that there are situations in which an individual must choose morally in order to choose rationally. For Harsanyi there is a rational way of choosing

[8] See Rawls, p. 12.

[9] See ibid., p. 11; 'the principles . . . are to assign basic rights and duties and to determine the division of social benefits.' Principles for individuals are distinguished from the principles of justice; see p. 108.

[10] See Rawls's distinction of 'the Reasonable' and 'the Rational', in 'Kantian Constructivism in Moral Theory', *Journal of Philosophy* 77 (1980), pp. 528–30.

[11] See Rawls, *A Theory of Justice*, pp. 14–15, and Harsanyi, 'Morality and the theory of rational behaviour', pp. 44–6, 56–60.

[12] See Harsanyi, 'Morality and the theory of rational behaviour', p. 62.

morally but no rational requirement to choose morally. And so again there is a basic difference between our theory and his.

Putting now to one side the views of Rawls and Harsanyi—views to which we shall often return in later chapters—we may summarize the import of the differences we have sketched. Our theory must generate, strictly as rational principles for choice, and so without introducing prior moral assumptions, constraints on the pursuit of individual interest or advantage that, being impartial, satisfy the traditional understanding of morality. We do not assume that there must be such impartial and rational constraints. We do not even assume that there must be rational constraints, whether impartial or not. We claim to demonstrate that there are rational constraints, and that these constraints are impartial. We then identify morality with these demonstrated constraints, but whether their content corresponds to that of conventional moral principles is a further question, which we shall not examine in detail. No doubt there will be differences, perhaps significant, between the impartial and rational constraints supported by our argument, and the morality learned from parents and peers, priests and teachers. But our concern is to validate the conception of morality as a set of rational, impartial constraints on the pursuit of individual interest, not to defend any particular moral code. And our concern, once again, is to do this without incorporating into the premisses of our argument any of the moral conceptions that emerge in our conclusions.

2.3 To seek to establish the rationality of moral constraints is not in itself a novel enterprise, and its antecedents are more venerable than the endeavour to develop moral theory as part of the theory of rational choice. But those who have engaged in it have typically appealed to a conception of practical rationality, deriving from Kant, quite different from ours.[13] In effect, their understanding of reason already includes the moral dimension of impartiality that we seek to generate.

Let us suppose it agreed that there is a connection between reason and interest—or advantage, benefit, preference, satisfaction, or individual utility, since the differences among these, important in other contexts, do not affect the present discussion. Let it further be agreed that in so far as the interests of others are not affected, a

[13] This conception of practical rationality appears with particular clarity in T. Nagel, *The Possibility of Altruism* (Oxford, 1970), esp. ch. x. It can also be found in the moral theory of R. M. Hare; see *Moral Thinking* (Oxford, 1981), esp. chs. 5 and 6.

person acts rationally if and only if she seeks her greatest interest or benefit. This might be denied by some, but we wish here to isolate the essential difference between the opposed conceptions of practical rationality. And this appears when we consider rational action in which the interests of others are involved. Proponents of the *maximizing* conception of rationality, which we endorse, insist that essentially nothing is changed; the rational person still seeks the greatest satisfaction of her own interests. On the other hand, proponents of what we shall call the *universalistic* conception of rationality insist that what makes it rational to satisfy an interest does not depend on whose interest it is. Thus the rational person seeks to satisfy all interests. Whether she is a utilitarian, aiming at the greatest happiness of the greatest number, or whether she takes into independent consideration the fair distribution of benefit among persons, is of no importance to the present discussion.

To avoid possible misunderstanding, note that neither conception of rationality requires that practical reasons be self-interested. On the maximizing conception it is not interests in the self, that take oneself as object, but interests of the self, held by oneself as subject, that provide the basis for rational choice and action. On the universalistic conception it is not interests in anyone, that take any person as object, but interests of anyone, held by some person as subject, that provide the basis for rational choice and action. If I have a direct interest in your welfare, then on either conception I have reason to promote your welfare. But your interest in your welfare affords me such reason only given the universalistic conception.

Morality, we have insisted, is traditionally understood to involve an impartial constraint on the pursuit of individual interest. The justification of such a constraint poses no problem to the proponents of universalistic rationality. The rational requirement that all interests be satisfied to the fullest extent possible directly constrains each person in the pursuit of her own interests. The precise formulation of the constraint will of course depend on the way in which interests are to be satisfied, but the basic rationale is sufficiently clear.

The main task of our moral theory—the generation of moral constraints as rational—is thus easily accomplished by proponents of the universalistic conception of practical reason. For them the relation between reason and morals is clear. Their task is to defend their conception of rationality, since the maximizing and

universalistic conceptions do not rest on equal footings. The maximizing conception possesses the virtue, among conceptions, of weakness. Any consideration affording one a reason for acting on the maximizing conception, also affords one such a reason on the universalistic conception. But the converse does not hold. On the universalistic conception all persons have in effect the same basis for rational choice—the interests of all—and this assumption, of the impersonality or impartiality of reason, demands defence.

Furthermore, and perhaps of greater importance, the maximizing conception of rationality is almost universally accepted and employed in the social sciences.[14] As we have noted, it lies at the core of economic theory, and is generalized in decision and game theory. Its lesser prominence in political, sociological, and psychological theory reflects more the lesser concern with rationality among many practitioners of those disciplines, than adherence to an alternative conception. Social scientists may no doubt be mistaken, but we take the onus of proof to fall on those who would defend universalistic rationality.

In developing moral theory within rational choice we thus embrace the weaker and more widely accepted of the two conceptions of rationality that we have distinguished. Of course, we must not suppose that the moral principles we generate will be identical with those that would be derived on the universalistic conception. Its proponents may insist that their account of the connection between reason and morals is correct, even if they come to agree that a form of morality may be grounded in maximizing rationality. But we may suggest, without here defending our suggestion, that few persons would embrace the universalistic conception of practical reason did they not think it necessary to the defence of any form of rational morality. Hence the most effective rebuttal of their position may be, not to seek to undermine their elaborate and ingenious arguments, but to construct an alternative account of a rational morality grounded in the weaker assumptions of the theory of rational choice.

3.1 Morals by agreement begin from an initial presumption against morality, as a constraint on each person's pursuit of his own

[14] See Harsanyi, 'Advances in Understanding Rational Behavior', p. 89; also J. Elster, *Ulysses and the Sirens: Studies in rationality and irrationality* (Cambridge, 1979); 'The "rational-choice" approach to human behaviour is without much doubt the best available model . . .', p. 112.

interest. A person is conceived as an independent centre of activity, endeavouring to direct his capacities and resources to the fulfilment of his interests. He considers what he can do, but initially draws no distinction between what he may and may not do. How then does he come to acknowledge the distinction? How does a person come to recognize a moral dimension to choice, if morality is not initially present?

Morals by agreement offer a contractarian rationale for distinguishing what one may and may not do. Moral principles are introduced as the objects of fully voluntary *ex ante* agreement among rational persons. Such agreement is hypothetical, in supposing a pre-moral context for the adoption of moral rules and practices. But the parties to agreement are real, determinate individuals, distinguished by their capacities, situations, and concerns. In so far as they would agree to constraints on their choices, restraining their pursuit of their own interests, they acknowledge a distinction between what they may and may not do. As rational persons understanding the structure of their interaction, they recognize a place for mutual constraint, and so for a moral dimension in their affairs.

That there is a contractarian rationale for morality must of course be shown. That is the task of our theory. Here our immediate concern is to relate the idea of such a rationale to the introduction of fundamental moral distinctions. This is not a magical process. Morality does not emerge as the rabbit from the empty hat. Rather, as we shall argue, it emerges quite simply from the application of the maximizing conception of rationality to certain structures of interaction. Agreed mutual constraint is the rational response to these structures. Reason overrides the presumption against morality.

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

Historically, moral contractarianism seems to have originated

among the Greek Sophists. Glaucon sketched a contractarian account of the origin of justice in Plato's *Republic* but significantly, he offered this view for Socrates to refute, not to defend.[15] Our theory of morals falls in an unpopular tradition, as the identity of its greatest advocate, Thomas Hobbes, will confirm. Hobbes transformed the laws of nature, which lay at the core of Stoic and medieval Christian moral thought, into precepts of reason that require each person, acting in his own interest, to give up some portion of the liberty with which he seeks his own survival and well-being, provided others do likewise.[16] But this agreement gives rise to actual constraint only through the efficacy of the political sovereign; from the standpoint of moral theory, the crucial step requires the intervention of a *deus ex machina*. Nevertheless, in Hobbes we find the true ancestor of the theory of morality that we shall present. Only recently has his position begun to acquire a significant following. G. R. Grice has developed an explicitly contractarian theory, and Kurt Baier has acknowledged the Hobbesian roots of his central thesis, that 'The very *raison d'être* of a morality is to yield reasons which overrule the reasons of self-interest in those cases when everyone's following self-interest would be harmful to everyone.'[17]

To the conceptual underpinning that may be found in Hobbes, Grice, and Baier, we seek to add the rigour of rational choice. Of course the resulting moral theory need not be one that they would endorse. But the appeal to rational choice enables us to state, with new clarity and precision, why rational persons would agree *ex ante* to constraining principles, what general characteristics these principles must have as objects of rational agreement, and why rational persons would comply *ex post* with the agreed constraints.

3.2 A useful vantage point for appreciating the rationale of constraint results from juxtaposing two ideas formulated by John Rawls. A contractarian views society as 'a cooperative venture for mutual advantage' among persons 'conceived as not taking an interest in one another's interests'.[18] The contractarian does not

[15] See Plato, *Republic*, 358b–359b.

[16] Thomas Hobbes, *Leviathan* (London, 1651), ch. 14, pp. 64–5.

[17] See G. R. Grice, *The Grounds of Moral Judgement* (Cambridge, 1967), and K. Baier, *The Moral Point of View: A Rational Basis of Ethics* (Ithaca, NY, 1958); the quotation is from p. 309.

[18] Rawls, *A Theory of Justice*, pp. 4, 13.

claim that all actual societies are co-operative ventures; he need not claim that all afford the expectation of mutual advantage. Rather, he supposes that it is in general possible for a society, analysed as a set of institutions, practices, and relationships, to afford each person greater benefit than she could expect in a non-social 'state of nature', and that only such a society could command the willing allegiance of every rational individual. The contractarian need not claim that actual persons take no interest in their fellows; indeed, we suppose that some degree of sociability is characteristic of human beings. But the contractarian sees sociability as enriching human life; for him, it becomes a source of exploitation if it induces persons to acquiesce in institutions and practices that but for their fellow-feelings would be costly to them. Feminist thought has surely made this, perhaps the core form of human exploitation, clear to us. Thus the contractarian insists that a society could not command the willing allegiance of a rational person if, without appealing to her feelings for others, it afforded her no expectation of net benefit.

If social institutions and practices can benefit all, then some set of social arrangements should be acceptable to all as a co-operative venture. Each person's concern to fulfil her own interests should ensure her willingness to join her fellows in a venture assuring her an expectation of increased fulfilment. She may of course reject some proposed venture as insufficiently advantageous to her when she considers both the distribution of benefits that it affords, and the availability of alternatives. Affording mutual advantage is a necessary condition for the acceptability of a set of social arrangements as a co-operative venture, not a sufficient condition. But we suppose that some set affording mutual advantage will also be mutually acceptable; a contractarian theory must set out conditions for sufficiency.

The rationale for agreement on society as a co-operative venture may seem unproblematic. The step from hypothetical agreement *ex ante* on a set of social arrangements to *ex post* adherence to those arrangements may seem straightforward. If one would willingly have joined the venture, why would one not now continue with it? Why is there need for constraint?

The institutions and practices of society play a co-ordinative role. Let us say, without attempting a precise definition, that a practice is co-ordinative if each person prefers to conform to it provided (most) others do, but prefers not to conform to it provided (most) others do

not.[19] And let us say that a practice is beneficially co-ordinative if each person prefers that others conform to it rather than conform to no practice, and does not (strongly) prefer that others conform to some alternative practice. Hume's example, of two persons rowing a boat that neither can row alone, is a very simple example of a beneficially co-ordinative practice.[20] Each prefers to row if the other rows, and not to row if the other does not. And each prefers the other to row than to act in some alternative way.

It is worth noting that a co-ordinative practice need not be beneficial. Among peaceable persons, who regard weapons only as instruments of defence, each may prefer to be armed provided (most) others are, and not armed provided (most) others are not. Being armed is a co-ordinative practice but not a beneficial one; each prefers others not to be armed.

The co-ordinative advantages of society are not to be underestimated. But not all beneficial social practices are co-ordinative. Let us say that a practice is beneficial if each person prefers that (almost) everyone conform to it rather than that (most) persons conform to no practice, and does not (strongly) prefer that (almost) everyone conform to some alternative practice. Yet it may be the case that each person prefers not to conform to the practice if (most) others do. In a community in which tax funds are spent reasonably wisely, each person may prefer that almost everyone pay taxes rather than not, and yet may prefer not to pay taxes herself whatever others do. For the payments each person makes contribute negligibly to the benefits she receives. In such a community persons will pay taxes voluntarily only if each accepts some constraint on her pursuit of individual interest; otherwise, each will pay taxes only if coerced, whether by public opinion or by public authority.

The rationale for agreement on society as a co-operative venture may still seem unproblematic. But the step from hypothetical agreement *ex ante* on a set of social arrangements to *ex post* adherence may no longer seem straightforward. We see why one might willingly join the venture, yet not willingly continue with it. Each joins in the hope of benefiting from the adherence of others, but fails to adhere in the hope of benefiting from her own defection.

In the next two chapters we shall offer an account of value,

[19] The discussion here is related to my characterization of a convention in 'David Hume, Contractarian', *Philosophical Review* 88 (1979), pp. 5–8.

[20] See Hume, *Treatise*, iii. ii. ii, p. 490.

rationality, and interaction, that will give us a precise formulation of the issue just identified. Prior to reflection, we might suppose that were each person to choose her best course of action, the outcome would be mutually as advantageous as possible. As we fill in our tax forms we may be reminded, *inter alia*, that individual benefit and mutual advantage frequently prove at odds. Our theory develops the implications of this reminder, beginning by locating the conflict between individual benefit and mutual advantage within the framework of rational choice.

3.3 Although a successful contractarian theory defeats the presumption against morality arising from its conception of rational, independent individuals, yet it should take the presumption seriously. The first conception central to our theory is therefore that of a morally free zone, a context within which the constraints of morality would have no place.[21] The free zone proves to be that habitat familiar to economists, the perfectly competitive market. Such a market is of course an idealization; how far it can be realized in human society is an empirical question beyond the scope of our enquiry. Our argument is that in a perfectly competitive market, mutual advantage is assured by the unconstrained activity of each individual in pursuit of her own greatest satisfaction, so that there is no place, rationally, for constraint. Furthermore, since in the market each person enjoys the same freedom in her choices and actions that she would have in isolation from her fellows, and since the market outcome reflects the exercise of each person's freedom, there is no basis for finding any partiality in the market's operations. Thus there is also no place, morally, for constraint. The market exemplifies an ideal of interaction among persons who, taking no interest in each other's interests, need only follow the dictates of their own individual interests to participate effectively in a venture for mutual advantage. We do not speak of a *co-operative* venture, reserving that label for enterprises that lack the natural harmony of each with all assured by the structure of market interaction.

The perfectly competitive market is thus a foil against which morality appears more clearly. Were the world such a market, morals would be unnecessary. But this is not to denigrate the value of morality, which makes possible an artificial harmony where

[21] This is the theme of ch. IV, below. See also my earlier discussion in 'No Need for Morality: The Case of the Competitive Market', *Philosophic Exchange* 3, no. 3 (1982), pp. 41–54.

natural harmony is not to be had. Market and morals share the non-coercive reconciliation of individual interest with mutual benefit.

Where mutual benefit requires individual constraint, this reconciliation is achieved through rational agreement. As we have noted, a necessary condition of such agreement is that its outcome be mutually advantageous; our task is to provide a sufficient condition. This problem is addressed in a part of the theory of games, the theory of rational bargaining, and divides into two issues.[22] The first is the bargaining problem proper, which in its general form is to select a specific outcome, given a range of mutually advantageous possibilities, and an initial bargaining position. The second is then to determine the initial bargaining position. Treatment of these issues has yet to reach consensus, so that we shall develop our own theory of bargaining.

Solving the bargaining problem yields a principle that governs both the process and the content of rationale agreement. We shall address this in Chapter V, where we introduce a measure of each person's stake in a bargain—the difference between the least he might accept in place of no agreement, and the most he might receive in place of being excluded by others from agreement. And we shall argue that the equal rationality of the bargainers leads to the requirement that the greatest concession, measured as a proportion of the conceder's stake, be as small as possible. We formulate this as the principle of minimax relative concession. And this is equivalent to the requirement that the least relative benefit, measured again as a proportion of one's stake, be as great as possible. So we formulate an equivalent principle of maximin relative benefit, which we claim captures the ideas of fairness and impartiality in a bargaining situation, and so serves as the basis of justice. Minimax relative concession, or maximin relative benefit, is thus the second conception central to our theory.

If society is to be a co-operative venture for mutual advantage, then its institutions and practices must satisfy, or nearly satisfy, this principle. For if our theory of bargaining is correct, then minimax relative concession governs the *ex ante* agreement that underlies a fair and rationale co-operative venture. But in so far as the social arrangements constrain our actual *ex post* choices, the question of compliance demands attention. Let it be ever so rational to agree to

[22] For references to the literature on rational bargaining, see notes 12–14 to ch. V, below.

practices that ensure maximin relative benefit; yet is it not also rational to ignore these practices should it serve one's interest to do so? Is it rational to internalize moral principles in one's choices, or only to acquiesce in them in so far as one's interests are held in check by external, coercive constraints? The weakness of traditional contractarian theory has been its inability to show the rationality of compliance.

Here we introduce the third conception central to our theory, constrained maximization. We distinguish the person who is disposed straightforwardly to maximize her satisfaction, or fulfil her interest, in the particular choices she makes, from the person who is disposed to comply with mutually advantageous moral constraints, provided he expects similar compliance from others. The latter is a constrained maximizer. And constrained maximizers, interacting one with another, enjoy opportunities for co-operation which others lack. Of course, constrained maximizers sometimes lose by being disposed to compliance, for they may act co-operatively in the mistaken expectation of reciprocity from others who instead benefit at their expense. Nevertheless, we shall show that under plausible conditions, the net advantage that constrained maximizers reap from co-operation exceeds the exploitative benefits that others may expect. From this we conclude that it is rational to be disposed to constrain maximizing behaviour by internalizing moral principles to govern one's choices. The contractarian is able to show that it is irrational to admit appeals to interest against compliance with those duties founded on mutual advantage.[23]

But compliance is rationally grounded only within the framework of a fully co-operative venture, in which each participant willingly interacts with her fellows. And this leads us back to the second issue addressed in bargaining theory—the initial bargaining position. If persons are willingly to comply with the agreement that determines what each takes from the bargaining table, then they must find initially acceptable what each brings to the table. And if what some bring to the table includes the fruits of prior interaction forced on their fellows, then this initial acceptability will be lacking. If you seize the products of my labour and then say 'Let's make a deal', I may be compelled to accept, but I will not voluntarily comply.

[23] This conclusion rests on a reinterpretation of the maximizing conception of rationality, which we develop in ch. VI, below; see especially the opening paragraph of 3. 1.

We are therefore led to constrain the initial bargaining position, through a proviso that prohibits bettering one's position through interaction worsening the position of another.[24] No person should be worse off in the initial bargaining position than she would be in a non-social context of no interaction. The proviso thus constrains the base from which each person's stake in agreement, and so her relative concession and benefit, are measured. We shall show that it induces a structure of personal and property rights, which are basic to rationally and morally acceptable social arrangements.

The proviso is the fourth of the core conceptions of our theory. Although a part of morals by agreement, it is not the product of rational agreement. Rather, it is a condition that must be accepted by each person for such agreement to be possible. Among beings, however rational, who may not hope to engage one another in a co-operative venture for mutual advantage, the proviso would have no force. Our theory denies any place to rational constraint, and so to morality, outside the context of mutual benefit. A contractarian account of morals has no place for duties that are strictly redistributive in their effects, transferring but not increasing benefits, or duties that do not assume reciprocity from other persons. Such duties would be neither rationally based, nor supported by considerations of impartiality.

To the four core conceptions whose role we have sketched, we add a fifth—the Archimedean point, from which an individual can move the moral world.[25] To confer this moral power, the Archimedean point must be one of assured impartiality—the position sought by John Rawls behind the 'veil of ignorance'. We shall conclude the exposition of our moral theory in Chapter VIII by relating the choice of a person occupying the Archimedean point to the other core ideas. We shall show that Archimedean choice is properly conceived, not as a limiting case of individual decision under uncertainty, but rather as a limiting case of bargaining. And we shall then show how each of our core ideas—the proviso against bettering oneself through worsening others, the morally free zone afforded by the perfectly competitive market, the principle of minimax relative concession, and the disposition to constrained maximization—may be related, directly or indirectly, to Archimedean choice. In embracing these other conceptions central to our theory, the Archimedean point reveals the coherence of morals by agreement.

[24] For the idea of the proviso, see note 1 to ch. VII, below.

[25] For the idea of an Archimedean point, see Rawls, *A Theory of Justice*, pp. 260–5.

4 A contractarian theory of morals, developed as part of the theory of rational choice, has evident strengths. It enables us to demonstrate the rationality of impartial constraints on the pursuit of individual interest to persons who may take no interest in others' interests. Morality is thus given a sure grounding in a weak and widely accepted conception of practical rationality. No alternative account of morality accomplishes this. Those who claim that moral principles are objects of rational choice in special circumstances fail to establish the rationality of actual compliance with these principles. Those who claim to establish the rationality of such compliance appeal to a strong and controversial conception of reason that seems to incorporate prior moral suppositions. No alternative account generates morals, as a rational constraint on choice and action, from a non-moral, or morally neutral, base.

But the strengths of a contractarian theory may seem to be accompanied by grave weaknesses. We have already noted that for a contractarian, morality requires a context of mutual benefit. John Locke held that 'an Hobbist . . . will not easily admit a great many plain duties of morality'.[26] And this may seem equally to apply to the Hobbist's modern-day successor. Our theory does not assume any fundamental concern with impartiality, but only a concern derivative from the benefits of agreement, and those benefits are determined by the effects that each person can have on the interests of her fellows. Only beings whose physical and mental capacities are either roughly equal or mutually complementary can expect to find co-operation beneficial to all. Humans benefit from their interaction with horses, but they do not co-operate with horses and may not benefit them. Among unequals, one party may benefit most by coercing the other, and on our theory would have no reason to refrain. We may condemn all coercive relationships, but only within the context of mutual benefit can our condemnation appeal to a rationally grounded morality.

Moral relationships among the participants in a co-operative venture for mutual advantage have a firm basis in the rationality of the participants. And it has been plausible to represent the society that has emerged in western Europe and America in recent centuries

[26] Locke MS, quoted in J. Dunn, *The Political Thought of John Locke* (Cambridge, 1969), pp. 218–19.

as such a venture. For Western society has discovered how to harness the efforts of the individual, working for his own good, in the cause of ever-increasing mutual benefit.[27] Not only an explosion in the quantity of material goods and in the numbers of persons, but, more important, an unprecedented rise in the average life span, and a previously unimaginable broadening of the range of occupations and activities effectively accessible to most individuals on the basis of their desires and talents, have resulted from this discovery.[28] With personal gain linked to social advance, the individual has been progressively freed from the coercive bonds, mediated through custom and education, law and religion, that have characterized earlier societies. But in unleashing the individual, perhaps too much credit has been given to the efficacy of market-like institutions, and too little attention paid to the need for co-operative interaction requiring limited but real constraint.[29] Morals by agreement then express the real concern each of us has in maintaining the conditions in which society can be a co-operative venture.

But if Locke's criticism of the scope of contractarian morality has been bypassed by circumstances that have enabled persons to regard one another as contributing partners to a joint enterprise, changed circumstances may bring it once more to the fore. From a technology that made it possible for an ever-increasing proportion of persons to increase the average level of well-being, our society is passing to a technology, best exemplified by developments in medicine, that make possible an ever-increasing transfer of benefits to persons who decrease that average.[30] Such persons are not party to the moral relationships grounded by a contractarian theory.

[27] We offer no explanation of this discovery. There seems no reason to suppose that it resulted from deliberate search.

[28] For the increase in average life span, see N. Eberstadt, 'The Health Crisis in the U.S.S.R.', *New York Review of Books* 28, no. 2 (1981), p. 23. For the broadening in the range of accessible occupations, note that 'As late as 1815 three-quarters of its [Europe's] population were employed on the land . . .', *The Times Concise Atlas of World History*, ed. G. Barraclough (London, 1982), p. 82.

[29] Thus the idea of economic man as an unlimited appropriator comes to dominate social thought. The effects of this conception are one of the themes of my 'The Social Contract as Ideology', *Philosophy and Public Affairs* 6 (1977), pp. 130–64.

[30] The problem here is not care of the aged, who have paid for their benefits by earlier productive activity. Life-extending therapies do, however, have an ominous redistributive potential. The primary problem is care for the handicapped. Speaking euphemistically of enabling them to live productive lives, when the services required exceed any possible products, conceals an issue which, understandably, no one wants to face. Without focusing primarily on these issues, I endeavour to begin a contractarian treatment of certain health care issues in 'Unequal Need: A Problem of Equity in Access to Health Care', *Securing Access to Health Care: The Ethical Implications of Differences in the Availability of Health Services*, 3 vols., President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research (Washington, 1983), vol. 2, pp. 179–205.

Beyond concern about the scope of moral relationships is the question of their place in an ideal human life. Glaucon asked Socrates to refute a contractarian account of justice, because he believed that such an account must treat justice as instrumentally valuable for persons who are mutually dependent, but intrinsically disvaluable, so that it 'seems to belong to the form of drudgery'.[31] Co-operation is a second-best form of interaction, requiring concessions and constraints that each person would prefer to avoid. Indeed, each has the secret hope that she can be successfully unjust, and easily falls prey to that most dangerous vanity that persuades her that she is truly superior to her fellows, and so can safely ignore their interests in pursuing her own. As Glaucon said, he who 'is truly a man' would reject moral constraints.[32]

A contractarian theory does not contradict this view, since it leaves altogether open the content of human desires, but equally it does not require it. May we not rather suppose that human beings depend for their fulfilment on a network of social relationships whose very structure constantly tempts them to misuse it? The constraints of morality then serve to regulate valued social relationships that fail to be self-regulating. They constrain us in the interests of a shared ideal of sociability.

Co-operation may then seem a second-best form of interaction, not because it runs counter to our desires, but because each person would prefer a natural harmony in which she could fulfil herself without constraint. But a natural harmony could exist only if our preferences and capabilities dovetailed in ways that would preclude their free development. Natural harmony would require a higher level of artifice, a shaping of our natures in ways that, at least until genetic engineering is perfected, are not possible, and were they possible, would surely not be desirable. If human individuality is to bloom, then we must expect some degree of conflict among the aims and interests of persons rather than natural harmony. Market and morals tame this conflict, reconciling individuality with mutual benefit.

[31] *Republic*, 358a, trans. A. Bloom (New York, 1968), p. 36.

[32] ibid., 359b, p. 37.

We shall consider, in the last chapters of our enquiry, what can be said for this interpretation of the place of moral relationships in human life. To do so we shall remark on speculative matters that lead beyond and beneath the theory of rational choice. And we may find ourselves with an alternative reading of what we present as a theory of morals.[33] We seek to forge a link between the rationality of individual maximization and the morality of impartial constraint. Suppose that we have indeed found such a link. How shall we interpret this finding? Are our conceptions of rationality and morality, and so of the contractarian link between them, as we should like them to be, fixed points in the development of the conceptual framework that enables us to formulate permanent practical truths? Or are we contributing to the history of ideas of a particular society, in which peculiar circumstances have fostered an ideology of individuality and interaction that coheres with morals by agreement? Are we telling a story about ideas that will seem as strange to our descendants, as the Form of the Good and the Unmoved Mover do to us?

[33] The thoughts in this paragraph have been influenced by R. Rorty; see esp. 'Method and Morality', in Norma Haan, R. N. Bellah, P. Rabinow, and W. M. Sullivan (eds.), *Social Science as Moral Inquiry* (New York, 1983), pp. 155–76.

II

CHOICE: REASON AND VALUE

1 'There is nothing either good or bad, but thinking makes it so.'[1] But if things considered in themselves are neither good nor bad, if there is no realm of value existing independently of animate beings and their activities, then thought is not the activity that summons value into being. Hume reminds us, 'Reason is, and ought only to be the slave of the passions', and while Hume's dictum has been widely disputed, we shall defend it.[2] Desire, not thought, and volition, not cognition, are the springs of good and evil.

We might wish to bypass the great questions of value that have exercised Western philosophers from the time of Socrates and the Sophists. But answers to these questions are implicit in the theory of rational choice. Although the presentation of a full theory of value lies beyond the scope of our present enquiry, yet we shall find that in discussing the basic concepts of choice theory—preference, utility, maximization—we are led not only to outline a technical apparatus, but also to reflect on reason and value. In this chapter we shall consider both the apparatus and the assumptions needed to explain morality as arising from rational agreement.

But rational agreement itself, and its role in interaction, must wait to be introduced. Our present focus is on *parametric* choice, in which 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. Interaction involves *strategic* choice, in which the actor takes his behaviour to be but one variable among others, so that his choice must be responsive to his expectations of others' choices, while their choices are similarly responsive to their expectations.[3] Rational choice is well defined for parametric environments; we shall argue that morality enters in defining it for certain strategic environments. But this is to anticipate. How shall we understand parametric choice?

[1] Shakespeare, *Hamlet*, II. ii. 255–7.
[2] Hume, *Treatise*, ii. iii. iii, p. 415.
[3] For the distinction between parametric and strategic rationality, see J. Elster, *Ulysses and the Sirens: Studies in rationality and irrationality* (Cambridge, 1979), pp. 18–19, 117–23.

product of labour, then there will be an effective demand for more labour—demand will exceed supply. The Marxist is thus caught in a contradiction. The buyer of labour power is able to extract surplus value—to pay a wage lower than the price he receives for the product of labour—only if the supply of labour exceeds the demand. But if there is surplus value to extract, this creates a demand for labour in excess of the existing supply. Or to put the matter another way, if the supply of labour exceeds the demand for it, this can only be because the cost of producing labour exceeds the price that can be received for its product. So there can be surplus value only if supply exceeds demand but if supply exceeds demand there can be no surplus value.

The operation of the perfectly competitive market must bring about an equilibrium between the supply of labour power and the demand for it, if—as Marxists suppose—labour power is a commodity brought into existence by the market. At equilibrium, there can be no surplus value for the buyer to extract and hence no exploitation of the seller. Now this proves nothing about the operation of imperfectly competitive markets. It proves nothing about a situation in which labour power comes into being as a by-product of other activities in which people engage for reasons quite other than meeting the market demand for labour. This could give rise to a permanent excess of labour power from which the owner of the means of production might benefit. And this might raise questions about the propriety of factor endowments that put some persons in a position to benefit from this over-supply of labour.

The appearance of market exploitation may be brought about by the very real exploitation actually occasioned by non-market features of society, or by features, such as the initial distribution of factors, taken as given in market interaction. And this may help to explain why the Marxist argument has convinced so many people, even though it is in fact incoherent and misdirected. But if it is a misdirected argument, which has nothing to do with the operation of perfectly competitive markets, then the Marxist theory of surplus value and exploitation serves only as a smoke-screen, concealing the very real exploitation that actual social arrangements, including of course those in 'Marxist' societies, may embody.

V

CO-OPERATION: BARGAINING AND JUSTICE

1.1 Reason, which increases the costs of natural interaction among human beings, offers not only a remedy for the ills it creates, but also the prospect of new benefits achieved mutually, through co-operation. Where the invisible hand fails to direct each person, mindful only of her own gain, to promote the benefit of all, co-operation provides a visible hand. In this chapter we begin our examination of co-operation as the rational response to market failure.

Where market interaction, with its pre-established harmony between equilibrium and optimum, is beyond good and evil, and natural interaction, in the presence of free-riders and parasites, degenerates into force and fraud, co-operative interaction is the domain of justice. Justice is the disposition not to take advantage of one's fellows, not to seek free goods or to impose uncompensated costs, provided that one supposes others similarly disposed. We shall show that in satisfying the conditions of practical rationality, co-operation ensures the elimination of the free-ridership and parasitism endemic to our natural condition, so that we may identify justice with the rational disposition to co-operative behaviour. Thus we find ourselves in agreement with the most influential contemporary theorist of justice, John Rawls, when he says, 'The circumstances of justice may be described as the normal conditions under which human cooperation is both possible and necessary.'[1]

David Hume, whose account of the circumstances of justice is followed by Rawls, supposes that the need for co-operation arises from the conjunction of scarcity, characterizing our 'outward circumstances', and a bias in favour of the self, characterizing our 'natural temper'.[2] The mutual unconcern presupposed by the market is an extreme form of self-bias, although the structure of market interaction makes it an innocuous one. If nature were to provide in

[1] J. Rawls, *A Theory of Justice* (Cambridge, Mass., 1971), p. 126.
[2] Hume, *Treatise*, iii. ii. ii, pp. 486–8; see also *Enquiry*, iii. i, pp. 183–6.

abundance the goods needed to satisfy our desires, or if benevolence were to lead each person to regard her fellows' concerns as her own, there would be no free-riders or parasites to be restrained by the visible hand of co-operation. All would seek naturally to co-ordinate their actions for the common good, without putting forward opposed claims to the fruits of their endeavours, which justice must resolve.

But scarcity and self-bias are not sufficient warrant for co-operative interaction. Were the scarcity faced by each person not aggravated by the presence of her fellows, then however self-biased she might be, her activities would bear little relation to those of others, and neither conflict nor co-operation would result. And could this scarcity not be alleviated by joint activities, then the domain of justice would extend only to the avoidance of mutually destructive conflicts, and not to the co-operative provision of mutual benefits. Thus among the circumstances necessary to justice we must include the variability of scarcity. The sources of satisfaction and dissatisfaction are not in fixed supply, so that by appropriate interaction overall costs may be lessened, and overall benefits increased.

Yet strictly speaking, it is not the fact of variable supply, but awareness of the fact, which is important. This awareness has both its negative and positive effects. On the one hand we become aware of each other as competitors for scarce goods, and this awareness exacerbates our competition, increasing our costs. It is to this awareness that we refer when we say that reason adds to the disadvantages of nature. On the other hand we become aware of each other as potential co-operators in the production of an increased supply of goods, and this awareness enables us to realize new benefits. Thus we note also that reason offers advantages unobtainable in nature.

Different theorists vary in the emphasis which they place on these aspects of our awareness. Thomas Hobbes focuses on the effect of recognizing one's fellows as competitors.[3] He argues that given scarcity and total mutual unconcern, each must view all other persons as in at least potential competition for the goods that she needs for survival or for greater well-being. But this creates in each person an actual preference for dominating her fellows; if she is able to establish her dominance now, then she may expect to be more

[3] See Hobbes, *Leviathan*, ch. 13, pp. 60–2.

successful in any future struggle for scarce goods. In this way potential conflict is converted into actual hostility, leading to the war of every man against every man, a war from which everyone must expect to suffer, yet from which no one dare abstain. To avoid this unending war each must agree to constrain her behaviour, provided others similarly agree, to the end that all may live peaceably. Justice, for Hobbes, lies in adherence to such agreement.

Hobbes's account reveals with stark clarity the role of reason in adding to the costs of natural interaction. For it is only a reasoning being, seeking to maximize her utility, who will adopt a pre-emptive strategy in interacting with her fellows. Those who are unaware of the prospect of bettering themselves by imposing costs on others will not find themselves in the peculiarly unpleasant condition which Hobbes describes. But Hobbes's account equally reveals the role of reason in enabling us to overcome this condition, for reason suggests the 'convenient Articles of Peace', agreement to which leads us out of the natural condition of humankind and eventually into society.[4]

Little in Hobbes's argument suggests a more positive role for co-operation. We are aware of each other as competitors, and so we come to co-operate in order to avoid mutually destructive conflict, but we are less aware of each other as potential sources of mutual benefit. Indeed, the world described by Hobbes resembles in some respects that which is said (with what accuracy I know not) to obtain among those South Pacific islanders, the Dobu.[5] Since it is the idea and not the reality which concerns us, we may ignore anthropological data which may or may not confirm our story, and suppose that the Dobu believe that the world offers only a fixed supply of the goods they treasure—primarily yams. The more yams in my garden, the fewer in yours. There is no place in the Dobuan scheme of things for co-operation directed at an increase of benefits, for more yams cannot be grown. At most, the costs of Hobbesian conflict can be avoided or alleviated. For even if each person is bent on acquiring, by the appropriate magical devices, the yams in her neighbours' gardens, it may still be better for each if no one seeks to pre-empt her fellows by designs, magical or otherwise, on their lives.

The Dobuan world may have its sophisticated defenders. Indeed, if we consider interaction among social classes rather than indivi-

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

[5] See Ruth Benedict, *Patterns of Culture* (Boston and New York, 1934), pp. 139–40, 146–8.

duals, the Marxists are foremost among them, for they suppose that in a class-based society, the more yams for the bourgeoisie, the fewer for the proletariat. And this leads some Marxists, insufficiently mindful of Hobbes, to deny the possibility of co-operation and the relevance of justice in class-based societies. But we shall suppose that the Dobuan view of the world is false. And given variable supply we may be aware of others as potential co-operators in increased production. It is this awareness which is emphasized by Hume and Rawls.

But recognition of the possibilities of increased production implicit in the idea of variable supply is also the basis of the market. And Hume and Rawls are insufficiently mindful of the role of the market in limiting the need for co-operation. Market interaction takes place under conditions of variable supply among persons who are mutually unconcerned. But given perfect competition they have no use for co-operative interaction as a visible hand, since the optimality of the market outcome excludes any alternative that would reduce overall costs or increase total benefits. Thus among the circumstances of justice, in addition to awareness of variable scarcity and individual bias, we must include recognition of the presence of externalities. Justice is concerned with both the inefficiencies occasioned by the imposition of costs without the provision of corresponding benefits, and those occasioned by the failure to provide benefits due to the inability to recover costs. Since externalities presuppose variable supply, we may then say that the fundamental circumstances of justice, those features of the human situation that give rise to co-operation, are awareness of externalities in our environment, and awareness of self-bias in our character.

1.2 When the market fails, each person, seeking to maximize her utility given the strategies she expects others to choose, fails to maximize her utility given the utilities those others receive. The equilibrium outcome of mutually utility-maximizing responses is not optimal. The dilemma of strategic rationality, posed starkly in the structure of the Prisoner's Dilemma, comes to possess our interactions, turning them away from the ends we seek. It can be exorcized only by changing the mode of interaction. Heretofore we have supposed that each person forms expectations about the choices of others to which she responds in choosing her own strategy, but that she considers only the costs and benefits to herself in making that choice. Externalities, unchosen third-party costs and benefits, are

ignored. And so in many situations the outcome brought about by the independent choices of those interacting is sub-optimal. In order to take effective account of externalities, each person must choose her strategy to bring about a particular outcome determined by prior agreement among those interacting. This agreement, if rational, will ensure optimality. It may of course be implicit rather than explicit, an understanding or convention rather than a contract. But it is not a mere fiction, since it gives rise to a new mode of interaction, which we identify as co-operation. In nature each person faces a separate strategic problem which must be solved in choosing her strategy. As co-operators, all persons face a common strategic problem which must be solved to determine every individual's choice of strategy.

As a preliminary characterization, to be modified in the next subsection, let us say that in co-operative interaction the primary object of choice is a set of strategies, one for each person. If co-operation is voluntary, as we shall assume unless we explicitly state otherwise, then each individual participates in, or agrees to, this primary choice. Since the product of a set of strategies, one for each interacting person, is an outcome, we may also say that in co-operative interaction the object of agreement is an outcome, which then determines each person's choice of strategy. In this chapter we shall determine the conditions for rational co-operative choice, or rational agreement on an outcome.

We have already stated one such condition: the object of rational co-operative choice must be an optimal outcome. We noted in Chapter III that an outcome may be considered either as a product of strategies or as set of utilities. We should expect that if each person selects her strategy independently of the others, as in natural interaction or in the market, then the first of these conceptions would take precedence. But if each selects her strategy as a result of agreement with the others, then since each in reaching agreement will be concerned with her utility, not her strategy, we should expect that the second conception would come to the fore. Now this suggests that in non-co-operative interaction the core rationality property is equilibrium, whereas in co-operative interaction the core rationality property is optimality. The three conditions on rational interaction introduced in Chapter III are therefore to be read in one way for the non-co-operative interaction of nature and the market, and in another way for co-operative interaction. Condition A, that each person's choice must be a rational response to the choices she

expects the others to make, must be read as requiring a utility-maximizing response in non-co-operative interaction, and as requiring an optimizing response in co-operation. Or in other words, in the absence of agreement on an outcome or set of strategies, it is rational for each person to seek to maximize her utility given the strategies she expects the others to choose, whereas in the context of agreement it is rational for each to seek an optimal outcome given the agreed strategies of the others. The first part of this proposition is a direct corollary of the identification of practical rationality with individual utility-maximization, but the corollary may seem so direct that the second part of the proposition is refuted. Our task is to show that this is not so, that co-operation is a rational mode of interaction.

In the two preceding paragraphs we have identified two quite distinct problems. Suppose we are to agree on an outcome or set of strategies; under what conditions is our agreement rational? This is the first problem; it concerns what we may call the internal rationality of co-operation, the rational way of making a co-operative choice. We shall solve this problem by finding a principle that rationalizes agreement in the way that the principle of expected utility-maximization rationalizes individual choice.

Under what conditions is it rational to agree to an outcome or set of strategies and act on that agreement? This is the second problem; it concerns the external rationality of co-operation. We must show when it is rational to act co-operatively rather than non-co-operatively. In resolving the first problem we take co-operation for granted; in turning to the second problem, we call co-operation into question and demand its rationale as a mode of interaction. The problems are of course related, but in this chapter we shall be concerned with the second only in so far as it bears on the first. Our resolution of it will come in Chapters VI and VII.

A third problem concerns the morality of co-operation. Under what conditions is co-operative interaction fair or impartial? And here, as with the market, we may distinguish the operation of co-operative practices from their conditions. In this chapter we shall examine the impartiality of the operation of co-operative practices and institutions in the light of our account of the internal rationality of co-operation, asking whether the principle of rational co-operative choice is an impartial principle. Thus there is a significant parallel between the argument of this chapter and that of the preceding one; we must exhibit the rationality and impartiality of co-operative interaction just as we exhibited the rationality and impartiality of market interaction. But we shall find that co-operation does not create a morally free zone; rather it requires moral constraints on maximizing behaviour.

In the next subsection we shall clarify the characteristics of co-operative interaction, showing how it can extend the range of available strategies and of (expected) outcomes. In section 2 we discuss alternative approaches to determining the conditions that co-operative choice must satisfy to be rational. In section 3 we set out in detail our own answer; we develop a principle for rational agreement. And in section 4 we examine the impartiality of rational co-operation.

1.3 Ernie and Bert want to meet. Ernie would prefer that they meet at the library, Bert at the cinema. Each is indifferent as to where he is should they fail to meet. So that we may provide an interval measure of their preferences, let us suppose that Ernie is indifferent between meeting at the cinema, and a lottery with equal chances of meeting at the library and not meeting, and that Bert is indifferent between meeting at the library, and a lottery with an equal chance of meeting at the cinema and not meeting. Calculating their utilities in accordance with the method of II.3.2, we arrive at this matrix representation of their situation:

	Bert goes to the library	Bert goes to the cinema
Ernie goes to the library	1, 1/2	0, 0
Ernie goes to the cinema	0, 0	1/2, 1

There are exactly two optimal outcomes; both must go to the same place. These outcomes are also in equilibrium. Unfortunately, the symmetry of the situation makes it impossible for Ernie and Bert to co-ordinate on one of these outcomes using Zeuthen's principle, as Victor and Valerie were able to do in III.2.2. If each chooses to go to his favoured meeting place they will fail to meet. If each, baffled, decides where to go by tossing a coin (that is adopting the mixed strategy which assigns a probability of 1/2 to each of his possible actions), then the expected outcome will be a lottery with a probability of 1/4 that they meet at the library, 1/4 that they meet at the cinema, and 1/2 that they do not meet, and the expected utility to each will be 3/8.

But suppose Ernie and Bert adopt a co-operative approach to

their interaction. Then a whole new range of possibilities opens up. For they can make a single, joint strategy choice rather than independent strategy choices. And the symmetry of their situation singles out a unique joint strategy, assigning a probability of 1/2 to both going to the library and 1/2 to both going to the cinema. Instead of each tossing a coin, they toss but one coin, agreeing that both will abide by the result. The expected outcome is then a lottery with a probability of 1/2 that they meet at the library and 1/2 that they meet at the cinema, and the expected utility to each is 3/4. This outcome is optimal, as indeed is the outcome of any joint mixed strategy that consists of a lottery with both going to the library and both going to the cinema as the only prizes. But none of these strategies, and so none of these outcomes, was available to Ernie and Bert in non-co-operative interaction.

As this example illustrates, co-operation may extend the range of strategies and outcomes. (It does not always do so; Noreen and Norman (III.2.2) could not extend their prospects by co-operating.) Let us define a *joint pure strategy* as the product of the members of a set of pure strategies, one for each person involved in interaction. A joint pure strategy is then simply a possible outcome (where we exclude merely expected outcomes). A *joint mixed strategy* is then a lottery with joint pure strategies, or possible outcomes, as prizes. An expected outcome is then simply a joint mixed strategy, or a lottery over possible outcomes. In co-operative interaction every lottery over possible outcomes determines an expected outcome. This need not be the case in non-co-operative interaction, where the set of expected outcomes is a subset (sometimes proper, sometimes improper) of the set of lotteries over possible outcomes.

In extending the range of strategies from individual to joint, we have implicitly altered our previous characterization of co-operative interaction, as having a set of strategies, one for each person, as the object of agreement. We should now say that the primary object of choice or agreement is a joint strategy. Since a set of strategies, one for each person, may always be represented by a joint strategy, but not conversely, this new characterization broadens our initial conception of co-operation.

We may conveniently represent any co-operative interaction involving only two persons in graphic form. We let the horizontal axis represent the utility scale of one person and the vertical axis represent the utility scale of the other. Each possible outcome is

represented by taking its utilities as co-ordinates. The closed, convex figure obtained by joining the points representing the possible outcomes then represents the range of lotteries over possible outcomes, or expected outcomes. Each point on or within this figure corresponds to an expected outcome, and each expected outcome corresponds to such a point. We call this figure the *outcome-space*. The upper right bound of this space then represents the range of optimal outcomes. Figures 1 and 2 illustrate two of our examples—Ernie and Bert, and Jane and Brian from Chapter III. We also show in Fig. 2 the point representing the unique outcome in equilibrium; its non-optimality is evident.

Although graphic treatment becomes difficult, this mode of representation may be extended to co-operative interactions involving any number of persons by employing an n-dimensional utility space with one dimension for each person's utilities. The outcome-space remains a closed, convex figure—although in n-dimensions rather than two—and its upper bound, generally n-1-dimensional, represents the range of optimal outcomes.

Although we shall not pursue a mathematical treatment here, we may sketch the formal problem that we seek to solve informally in this chapter. We want to define a choice function, which we call the *co-operation function*. Its domain is a set of outcome-spaces. Its values are the members of a set of points in utility space. The idea, of

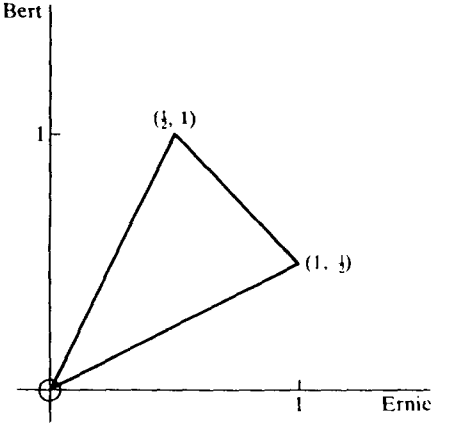

Figure 1

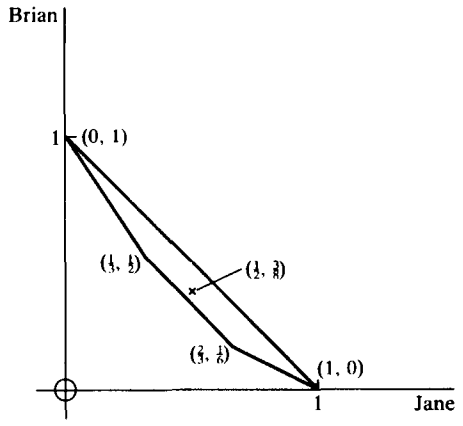

Figure 2

course, is that for an outcome-space, representing the outcomes of a possible co-operative interaction, the function determines a point, representing the particular outcome that determines the co-operators' joint strategy. One condition that we should want to impose on the co-operation function is that its domain be complete, including every possible outcome-space. A second condition is that for every member of its domain, the function determine a unique point as its value. We may specify this more closely; the point must be included in the outcome-space (so that it represents an outcome of the interaction) and indeed must fall on the upper bound of the outcome-space (so that it represents an optimal outcome). These are only necessary conditions, too weak to determine a unique function. We shall pursue the task of stating, although only informally, a sufficiently strong set of conditions in section 3. But we shall begin by considering three ways in which one might propose to arrive at a co-operation function.

2.1 Several persons are to agree on an outcome which is then to be brought about by a joint strategy determining each person's action. Under what conditions is their agreement rational? Individual choice is rational in so far as it is utility-maximizing. Is there an analogue to utility-maximization for agreement or co-operative choice? Since it is to be voluntary, it must reflect, and in some sense reflect equally, the preferences of each person. A first proposal,

therefore, would be that we derive, from the individual preference orderings of the co-operators, a social preference ordering. We then define a measure of social preference, and identify rational co-operative choice with the maximization of this measure.

The reader familiar with the literature of social choice will recognize the problem facing this proposal. Kenneth Arrow has demonstrated that it is not possible to derive a social preference ordering from every configuration of individual preference orderings without violating one of the following conditions:[6]

(1) If each individual prefers outcome X to outcome Y, then society prefers X to Y;
(2) There is no individual whose preferences are automatically society's preferences, whatever the preferences of the others;
(3) Social preference over any outcomes depends only on individual preferences over those outcomes.

Since we require that co-operative choice select an optimal outcome, we cannot reject condition (1). Since we require that co-operative choice reflect the preferences of each person, we cannot reject condition (2). And since on this proposal social preference is based on and only on individual preferences, we surely must require it to be based on relevant individual preferences, so that we cannot reject condition (3). Hence we cannot derive a social preference ordering for every configuration of individual preference orderings and the first suggestion would seem to be a non-starter.

But let us not proceed too hastily. We suppose that an interval measure, utility, can be defined over the preferences of each co-operator. A modified proposal, then, is that we derive, from these interval measures, a social interval measure, and identify rational co-operative choice with the maximization of this measure. However, this modified proposal does not enable us to escape from Arrow's demonstration. Recall that we have introduced no basis for comparing the utilities of different persons. Given interpersonal incomparability, Arrow's demonstration may be reproduced for the problem of defining a social interval measure on the basis of individual interval measures.[7]

A more promising modification is suggested by the formal state-

[6] See K. J. Arrow, 'Values and Collective Decision-Making', in P. Laslett and W. G. Runciman (eds.), *Philosophy, Politics and Society*, 3rd series (Oxford, 1967), pp. 225–6.

[7] See R. D. Luce and H. Raiffa, *Games and Decisions* (New York, 1957), pp. 344–5.

ment of our problem in the preceding section. We seek a function that determines an outcome for every outcome-space, or in effect, an outcome for every set of outcomes on the basis of the individual utilities assigned to the members of the set. Why not bypass a social preference ordering or social interval measure and go directly from the individual measures to the chosen outcome? This may seem to be a dodge, but it is a dodge that works; we may derive a social choice from every configuration of individual preference orderings and satisfy these conditions:

(1′) If each individual prefers outcome X to outcome Y, then society does not choose Y if X is available;
(2′) There is no individual whose utility is automatically maximized by society's choices, whatever the utilities of the others;
(3′) Social choice over any outcomes depends only on individual utilities for those outcomes.

The revised proposal, then, is that we derive, from the individual preference orderings of the co-operators, a social choice, by a function including all configurations of individual orderings in its domain and satisfying conditions (1′), (2′), and (3′). But the only plausible way to satisfy our conditions is to include, in the social choice, the entire range of optimal outcomes.[8] Rather than picking out a single outcome, we find ourselves with no basis for selecting among those that survived our previous statement of necessary conditions on co-operative choice.

But again, let us not proceed too hastily. In the literature of social choice, lotteries are generally ignored; outcomes are treated as possible outcomes, excluding expected outcomes that are lotteries with possible outcomes as prizes. Since we have no objection to lotteries, why should we not suppose that, if our function for determining social choice selects all optimal outcomes, then operationally it selects the lottery that assigns equal probability to all of the optimal possible outcomes, and which must itself be an expected outcome? There is an immediate problem with this new proposal; the expected outcome determined by a lottery that assigns equal probability to each of the optimal possible outcomes need not be

[8] See discussion in A. K. Sen, *Collective Choice and Social Welfare* (San Francisco, 1970), pp. 47–50. Our claim is stronger than anything defended by Sen, and relies in part on work in progress.

itself optimal.[9] Perhaps a further modification would accommodate this problem. But since our concern is not to examine formal issues in the theory of social choice, we shall not pursue this here. We have shown that, despite Arrow's notorious demonstration of the impossibility of deriving a social preference ordering from individual preference orderings, it may be feasible to relate co-operative choice to individual preferences in such a way that the choice in any situation depends only on the individual utilities for the outcomes of the situation.

But feasible as it may be, we reject the view that agreement or co-operative choice should rest only on individual utilities, because it leads to treating all optimal outcomes as rationally indifferent. This may be appropriate if we think of social choice as choice, by some person in authority, or some authoritative body, taking into account the preferences of those affected by the choice, and excluding all other considerations. Those affected are treated as passive. The outcome cannot be affected by any interaction among them. They are recipients or consumers of the goods (and, perhaps, bads) to be provided; they are not producers of those goods.

The context of co-operative choice is quite different. Although co-operators are concerned with the utility outputs rather than the strategic inputs of the outcomes among which they must choose, yet they are not passive in the process of agreement, and its strategic character may not be ignored in our analysis. Given a range of optimal outcomes, each person has a definite preference ordering over the members of the range, and since the outcomes are optimal, each preference of each person must conflict with some preference of some other person. As rational, each seeks to maximize her own utility, and so to exert herself to secure that outcome or those outcomes that she favours. To suppose that the optimal outcomes are to be taken as indifferent, so that selection among them proceeds simply by lot, is to treat the process of agreement quite apart from its actual dynamic character. And in this process some outcome or outcomes may be singled out in ways that do not depend directly and solely on the preference orderings of the individual co-opera-

[9] The proof is simple and proceeds by example. Consider a situation with three optimal possible outcomes, represented by the points (0.1, 1), (0.7, 0.7), and (1, 0.1), and a fourth possible outcome at (0, 0). A lottery assigning equal probabilities to each of the optimal points determines the expected outcome represented by the point (0.6, 0.6), which is evidently sub-optimal.

tors. Co-operative choice must reflect the preferences of each person, if it is to be rational, but how it reflects their preferences must depend on the structure of their interaction, on the consequences for everyone of what each is able to do. We shall show presently how several persons may choose an outcome to be brought about by a joint strategy, taking into account their status as actors. Social choice, in ignoring this status (quite reasonably, given its context of application) proves irrelevant to the co-operative choice of a joint strategy.

2.2 Defeated in its assault on the market, utilitarianism returns to the battle in offering itself as a guide to rational agreement.[10] We asked at the beginning of the preceding subsection whether there is an analogue to utility-maximization for co-operative choice. The utilitarian replies that there is indeed an analogue, *welfare-maximization*, where welfare is defined as the sum of individual utilities. The utilitarian differs in his approach from that considered in 2.1, in insisting on the possibility of an interpersonally comparable measure of individual utility, which enables us to sum the utilities of different persons, and thus determine social welfare. The utilitarian proposal identifies rational co-operative choice with the maximization of this sum.

Since co-operative choice assumes a fixed group of co-operators, the proposal that the sum of individual utilities be maximized is equivalent to the proposal that the *average* of individual utilities be maximized. (The average is simply the sum divided by the number of persons.) This enables the utilitarian to enlist the support of John Harsanyi, who has demonstrated that if social welfare meets the same rationality requirements as individual utility (if, in other words, it affords an interval measure defined over the outcomes), and if the welfare of an outcome is positively related to the individual utilities of that outcome, then welfare must be a weighted average of those utilities.[11] Furthermore, if the weighting is not

[10] The terminology is not his, but the view is that of Harsanyi. He claims that 'the updated version of classical utilitarianism is the only ethical theory which consistently abides by the principle that moral issues must be decided by rational tests and that moral behaviour itself is a special form of rational behaviour', 'Morality and the theory of rational behaviour', p. 40.

[11] The demonstration may be found in J. C. Harsanyi, 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', in *Essays on Ethics*, pp. 10–15; also in *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge, 1977), pp. 64–9.

arbitrary, but reflects a common measure of utility, then social welfare must be simply the average of individual utilities. The utilitarian may then claim that his proposal is the uniquely rational method of taking into account the preferences of each in choosing an outcome to be brought about by a joint strategy.

Superficially, the utilitarian proposal may seem very different from those examined in the preceding subsection. There we found that all optimal outcomes were rationally indifferent. The utilitarian provides a measure that distinguishes among them. But this contrast conceals a deeper affinity between utilitarianism and the approach of social choice. Both suppose that the choice of an outcome is to be based on and solely on the preferences of those affected. Both ignore the process of agreement among persons each of whom is concerned to maximize her own utility. The utilitarian proposal, applied to co-operation, treats the co-operators as passive recipients of goods, not as actively engaged in producing them and agreeing on their distribution. In their interaction, no one need take an interest in the outcome which maximizes average utility, except that particular no one who is the average person. But even should the average person be one of the co-operators, he is but one among the many who must participate in co-operative choice.

In examining the market we noted that the welfare optimum does not in general coincide with the optimal outcome achieved through perfect competition. The redistribution that would be required to bring the outcome of market interaction into accord with utilitarian requirements could not be viewed as rational by some of the persons in the market. We may suppose that similarly the welfare optimum would not in general coincide with the outcome achieved through co-operation. For the utilitarian ignores, as his principles require him to ignore, the structure of interaction. From the perspective of co-operation, utilitarianism as a proposal for choosing the outcome seems to have nothing to recommend it.

We have, of course, no quarrel with Harsanyi's demonstration; one does not quarrel with proofs. We agree that an interval measure defined over outcomes and positively based on the individual interval measures of preference must be utilitarian. If an interpersonally comparable measure of utility can be defined, then utilitarianism would seem to provide the rationally required method of social choice. But we deny the relevance of Harsanyi's demonstration to our problem. Rational co-operative choice must reflect the

preferences of the co-operators, but their preferences as actors. We turn without further ado to an approach to co-operative choice that takes interaction, as well as preference, fully into account.

2.3 Co-operation arises from the failure of market interaction to bring about an optimal outcome because of the presence of externalities. We may then think of co-operative interaction as a visible hand which supplants the invisible hand, in order to realize the same ideal as the market provides under conditions of perfect competition. In market exchange, costs and benefits are related in a manner that is not only optimal, but that affords each person a return equal in value to her marginal contribution. The joint strategy selected for co-operative interaction should bring about the same relation of costs and benefits under conditions in which the provision of a benefit may not directly result in the recovery of its costs, or the imposition of a cost may not directly require the provision of a compensating benefit. In accepting the joint co-operative strategy as the basis for her own actions, each individual forgoes the opportunities for free-ridership or parasitism that imperfect competition affords, in return for others forgoing their similar opportunities. But of course each endeavours to have the joint strategy chosen that is most favourable to herself, minimizing the costs of her restraint and maximizing the benefits she receives from the restraint of others. In reaching agreement on a joint strategy, then, each individual sees herself engaged in a process of *bargaining* with her fellows. Through bargaining, individuals arrive at a basis for co-operative interaction that enables them to relate costs and benefits, despite the presence of externalities, in the way automatically brought about by the market in the absence of those externalities.

The idea of a bargain enables us to incorporate into our account of rational co-operative choice what is missing from the perspectives of social choice and utilitarianism—the active involvement of the co-operators. It also enables us to capture the requirement that agreement on a joint strategy be voluntary. A rational bargain ensures the participation of each in reaching an agreed outcome. As we noted at the beginning of this chapter, not all co-operation is based on actual agreement. We may not then suppose that every joint strategy is chosen by a bargaining procedure in which all of those basing their actions on the strategy participated. But for co-operation to be rational, we must suppose that the joint strategy would have been chosen through such a procedure, so that each

person, recognizing this, may voluntarily accept the strategy. Each is then able to view the distribution of the benefits realized from co-operation as acceptable to her as an actor, a full participant in the co-operative process.

Before turning to our account of rational bargaining, we should guard against one possible misunderstanding of our argument. Co-operative interaction is not itself bargaining. Co-operative interaction results from, and is determined by, the choice of a joint strategy. Each then chooses her own actions as required by that strategy; in so doing she is not bargaining with her fellows. Rather, choosing this joint strategy involves bargaining. Bargaining thus gives rise to co-operative interaction but is itself non-co-operative. This distinction is of great importance in subsequent discussion, for as we shall see, in co-operating persons must at times constrain their utility-maximizing behaviour, but in bargaining itself persons accept no such constraint. The constraints required by co-operation are arrived at through bargaining, but are no part of the bargaining process. To refer back to what should be a familiar distinction at this point in our argument, in bargaining, each person's behaviour must be a utility-maximizing response to her expectations of others' behaviour, whereas in co-operating each person's behaviour must be an optimizing response to the expectations she forms of others' behaviour on the basis of their joint strategy.

3.1 The general theory of rational bargaining is underdeveloped territory. Whether there are principles of rational bargaining with the same context-free universality of application as the principle of expected utility maximization has been questioned, notably by Alvin Roth.[12] John Harsanyi insists that there is a general theory, and claims to have constructed it, building on earlier work of Frederik Zeuthen and John Nash.[13] Although the Zeuthen–Nash–Harsanyi approach has commanded widest support among those who accept the possibility of a general theory, it is not without competitors. Undaunted both by Roth's scepticism and by Harsanyi's

[12] See A. E. Roth, M. W. K. Malouf, and J. K. Murnighan, 'Sociological versus Strategic Factors in Bargaining', *Journal of Economic Behavior and Organization* 2 (1981), pp. 174–7.

[13] See F. Zeuthen, *Problems of Monopoly and Economic Welfare* (London, 1930), ch. 4; J. F. Nash, 'The Bargaining Problem', *Econometrica* 18 (1951), pp. 155–62, and 'Two-person Cooperative Games', *Econometrica* 21 (1953), pp. 128–40; Harsanyi, *Rational Behavior*.

dogmatism, we shall outline our own theory.[14] Interested readers will find our reasons for preferring it to the Zeuthen–Nash–Harsanyi theory in 3.4.

In any bargain it is necessary first to specify the initial position of the parties—the *initial bargaining position*, as we shall call it. We may think of this initial position as an outcome, or as a set of utilities, one for each bargainer. A bargaining situation then consists of a set of outcomes, any one of which may be realized by the bargainers through agreement on the appropriate joint strategy, and a specially designated outcome, the initial bargaining position. In 1.3 we introduced the representation of the set of outcomes graphically, as a closed, convex figure in utility space. The initial bargaining position is of course represented by a point in that space. Fig. 3 depicts a typical two-person bargaining situation.

The initial bargaining position fixes a base point from which bargaining proceeds. The utility it affords each person represents, in effect, what she brings to the bargaining table, and is not part of what she seeks to gain at the table from the bargain. In agreeing to a joint strategy the bargainers are concerned with the distribution of only the utility that each may receive over and above what she obtains in the initial bargaining position. We shall say that the bargainers are concerned with the distribution of the gains which co-operation may bring them, or with the co-operative *surplus*. The initial bargaining position identifies that part of each person's utility that is not part of the co-operative surplus.

In 1.3 we introduced the idea of a co-operation function. There we

[14] The underlying idea of our theory appears in 'Rational Co-operation', *Noûs* 8 (1974), pp. 53–65. A brief sketch appears in 'Economic Rationality and Moral Constraints', *Midwest Studies in Philosophy* 3 (1978), pp. 92–3, and a fuller account in 'The Social Contract: Individual Decision or Collective Bargain?', in C. A. Hooker, J. J. Leach and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, 2 vols. (Dordrecht, 1978), vol. 2, pp. 47–67. An informal account is found in 'Bargaining Our Way Into Morality: A Do-It-Yourself Primer', *Philosophic Exchange* 2, no. 5 (1979), pp. 14–27. More recently still, see 'Justified Inequality?', *Dialogue* 21 (1982), pp. 431–43, and 'Justice as Social Choice', in D. Copp and D. Zimmerman (eds.), *Morality, Reason and Truth: New Essays on the Foundations of Ethics* (Totowa, N J, 1985), pp. 251–69. For two-persons situations, our account yields solution G, discussed in A. E. Roth, *Axiomatic Models of Bargaining* (Berlin, 1979), pp. 98–108, and axiomatized by E. Kalai and M. Smorodinsky, 'Other Solutions to Nash's Bargaining Problem', *Econometrica* 43 (1975), pp. 513–18. But for situations involving more than two persons, our account departs from solution G in those cases in which G fails to be Pareto-optimal. See my paper 'Bargaining and Justice', appearing in *Social Philosophy and Policy* 2, no. 2.

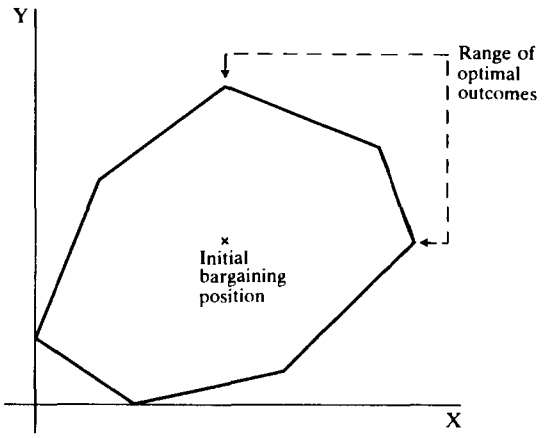

Figure 3

suggested that its domain was a set of outcome-spaces, each space representing the outcomes of a possible co-operative interaction. We must now revise that account to accommodate the initial bargaining position. The domain of the co-operation function is a set of (representations of) bargaining situations, that is, a set of outcome-spaces each with its specially designated point representing the initial bargaining position. For each bargaining situation, the co-operation function determines a point representing the outcome to be realized through co-operative interaction. In 1.3 we stated that this point must fall on the upper bound of the outcome-space, since it must represent an optimal outcome. We may now add that this point must represent an outcome affording each person a utility no less than her utility in the initial bargaining position. In Figure 4 we illustrate the effect of this requirement, in determining what we shall call the range of *admissible* optimal outcomes.

The inclusion of the initial position in the bargaining situation formally distinguishes our account of the co-operation function from those proposed on the basis of social choice or of utilitarianism. On both of these views, nothing but the preferences of those concerned (or the measure of those preferences) is relevant to the choice of an outcome. On our view, preferences with respect to a particular state of affairs are singled out for special consideration.

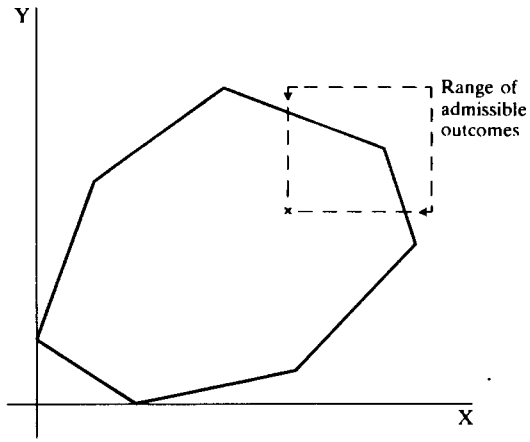

Figure 4

Neither social choice nor utilitarian treatments allow such preferences to restrict the range of optimal outcomes available for selection. And as we shall see, the initial bargaining position has further effects on the procedure of bargaining.

But how is the initial bargaining position to be determined? If we suppose co-operation to parallel market interaction in its structure, then we might propose to identify the initial position with the initial distribution of the factors to be used in co-operation, treating it as the set of pay-offs or utilities directly realizable from the initial factor endowments. But this identification would seem to ignore the place of co-operation in our argument. We have introduced it as an alternative to natural and market interaction, entered into because of their failure to provide optimal outcomes in the face of externalities. The co-operative surplus, we might then suppose, is what may be gained over and above what would result from non-co-operative interaction. Thus we might propose to identify the initial bargaining position with the non-co-operative outcome.

But non-co-operative interaction is marred by force and fraud, by the presence of free-riders and parasites. Co-operation is intended to eliminate all of these. Is it then rational for each bargainer to enter into agreement with her fellows if they may bring to the bargaining table whatever they might obtain from free-ridership and parasitism,

or more correctly the utility equivalent, and to consider the distribution only of the additional goods which co-operation provides?

We must here distinguish two quite different issues. On the one hand, we may ask what each person may bring to the bargaining table, if co-operation is to be rational. On the other hand, we may ask what each person must take from the bargaining table, if co-operation is to be rational. We may agree that each person must take from the bargain the expectation of a utility at least equal to what she would expect from non-co-operative interaction, if she is to find it rational to co-operate. It does not follow that she must bring such a utility to the bargain, as determining her share of the base point from which bargaining proceeds.

The determination of the initial bargaining position raises some of the most complex issues that we must examine. But we need not examine them here; indeed, we shall defer them to Chapter VII. They concern the external rationality of co-operation. The initial bargaining position must be fixed in such a way that it is rational for persons to enter into co-operation, agreeing to a joint strategy and acting on their agreement. Here our concern is with the internal rationality of co-operation, the rational way of choosing the joint strategy for interdependent interaction. Thus in this chapter we shall examine the initial bargaining position only in terms of its role in the selection of the co-operative strategy.

3.2 The procedure of bargaining may be divided into two principal stages. First, each party advances a claim—proposes an outcome or joint strategy for mutual acceptance. In general the claims of the parties are incompatible. Hence second, each party—or at least some party—offers a concession by withdrawing some portion of his original claim and proposing an alternative outcome. Barring deadlock the process of concession continues until a set of mutually compatible claims is reached. We may simplify this second stage if we suppose that after each party advances his initial claim, agreement is reached in a single round of concessions. What claims, and what concessions, will rational bargainers make?

Each person expects that what he gets will be related to what he claims. Each wants to get as much as possible; each therefore claims as much as possible. But in deciding how much is possible, each is constrained by the recognition that he must neither drive others away from the bargaining table, nor be excluded by them. Hence each person's claim is limited by the overall co-operative surplus,

and more specifically by the portion of the surplus that it is possible for him to receive. To claim more would be to propose that others give up some of what they brought to the bargaining table, some of their pay-off in the initial bargaining position. Since no rational person can expect any other rational person to do this, to claim more than one's largest possible portion of the co-operative surplus would be idle, or worse since if one were to press such a claim, one would only drive others away or face exclusion oneself. Since one wants to benefit from a share of the co-operative surplus, one has no interest in causing the process of bargaining to fail, as its failure would result in no co-operation and so no surplus. Each person, seeking to maximize his own utility and aware that others are seeking to maximize theirs, thus claims the outcome or joint strategy that would maximize the utility he can receive from the co-operative surplus. Or in other words, each person proposes, from among the admissible outcomes, the one that maximizes his utility. In Fig. 5 we show the claims that rational individuals would advance in several bargaining situations. Inspection makes it apparent that in two-person bargaining, claim points are easily determined from the outcome-space and the point representing the initial bargaining position.

However, we must beware lest consideration of two person bargaining lead us to misunderstand the determination of claims. In a situation involving more than two persons, each person may not always claim all of the co-operative surplus that he might receive, but only that part of the surplus to the production of which he would contribute. Each person's claim is bounded by the extent of his participation in co-operative interaction. For if someone were to press a claim to what would be brought about by the co-operative interaction of others, then those others would prefer to exclude him from agreement.

Given the claims of the bargainers, what concessions is it rational for them to make? To answer this question we must first consider how concessions are to be measured. The absolute magnitude of a concession, in terms of utility, is of course the difference between the utility one would expect from the outcome initially claimed and the utility one would expect from the outcome proposed as a concession. But this magnitude offers no basis for relating the concessions of different bargainers, since the measure of individual utility does not permit interpersonal comparisons. However, we may introduce a

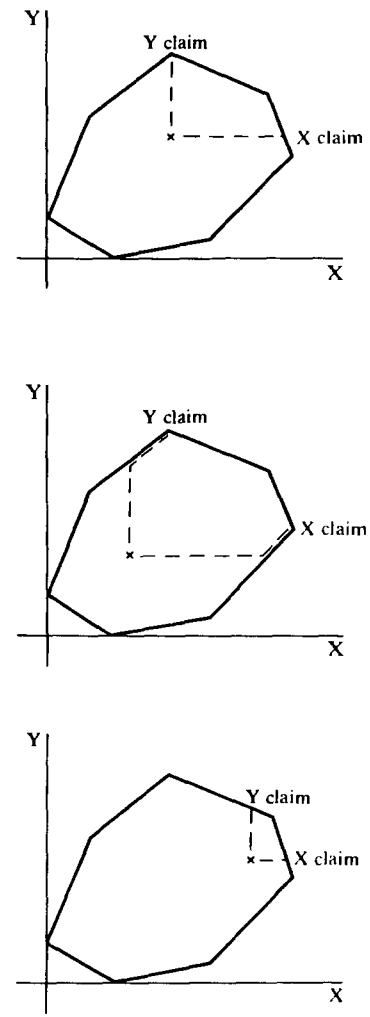

Figure 5

measure of *relative concession* which does enable us to compare the concessions of different bargainers, and which thus gives us a basis for determining what concession each must rationally make.

Suppose one were to receive one's claim. Then one would make no concession whatsoever. Suppose on the other hand one were to end

up in the initial bargaining position. Then one would in effect make a complete concession; one would receive no more than what one brought to bargaining, no more than what one had been allotted in the claims of the other bargainers. The *relative magnitude* of any concession may be expressed as the proportion its absolute magnitude bears to the absolute magnitude of a complete concession. If the initial bargaining position affords some person a utility u^*, and he claims an outcome affording him a utility u^*, then if he concedes an outcome affording him a utility u, the absolute magnitude of his concession is $(u^* - u)$, of complete concession $(u^* - u^*)$, and so the relative magnitude of his concession is $[u^* - u)/(u^* - u^*)]$.

Relative concession is a proportion of two utility-differences or intervals. Now an interval measure, such as utility, fixes these proportions independently of its arbitrary features—the zero-point and the unit. Consider three temperatures, as shown on both the Celsius and Fahrenheit scales: let them be 10 °C = 50 °F, 20 °C = 68 °F, and 25 °C = 77 °F. Measured on the Celsius scale, the difference or interval between the first and second is 10 °C, between the second and third 5 °C. Measured on the Fahrenheit scale, the intervals are 18 and 9 °F. But the proportion between the two intervals, 10/5 and 18/9, is 2, independent of the choice of scale. This illustrates for temperature what holds for all interval measures and so for utility. Relative concession is independent of the choice of utility scale. Each person's relative concessions are fixed no matter how we choose to measure his utilities.

Furthermore, we have introduced relative concession so that for each person, the relative magnitude of no concession is always 0, and the relative magnitude of complete or full concession is always 1. Thus we have a measure of relative concession which, without introducing any interpersonal comparison of utility (for we do not suppose that equal relative concessions have equal utility costs), nevertheless enables us to compare the concessions advanced by different persons in a bargaining situation.

We now represent each outcome in the admissible range, not in terms of the utilities it affords to each person, but in terms of the concessions that would be required from each, were the outcome to be chosen as determining the joint strategy for co-operative interaction. Zeuthen's principle may then be employed to provide a rule deciding what concessions must rationally be made and so what outcome must be accepted. It will be remembered from III.2.2 that

this principle states that the person whose ratio between cost of concession and cost of deadlock is less must concede to the other. Or in terms of our present discussion, the principle states that the person with a lesser relative concession must concede. Extending this rule to bargaining among several persons, we claim that the principle should state that given a range of outcomes, each of which requires concessions by some or all persons if it is to be selected, then an outcome be selected only if the greatest or *maximum* relative concession it requires, is as small as possible, or a *minimum*, that is, is no greater than the maximum relative concession required by every other outcome. We call this the principle of minimum-maximum, or *minimax relative concession.*

Let us return to a familiar example—Jane and Brian. Suppose that we identify the initial bargaining position with the unique equilibrium outcome they might expect to result from non-co-operative interaction; it affords Jane a utility of 1/2 and Brian a utility of 3/8. We illustrate this situation in Fig. 6. The admissible outcomes are those optimal outcomes affording Jane a utility of at least 1/2 and Brian a utility of at least 3/8; the points representing them fall on the line joining (0, 1) and (1, 0), between the points (1/2, 1/2) and (5/8, 3/8) inclusive. It is evident that Jane claims the latter and Brian the former. The outcome with utility to Jane 9/16 and utility to Brian 7/16 requires from Jane a relative concession of $[(5/8-9/16)/(5/8-1/2)]=1/2$, and from Brian a relative concession of $[(1/2-7/16)/(1/2-3/8)]=1/2$. Any admissible outcome more favourable to Jane will clearly demand a larger relative concession from Brian and any outcome more favourable to Brian will demand a larger relative concession from Jane. Hence by the principle of minimax relative concession, Jane and Brian should agree on this outcome, affording her 9/16 and him 7/16; if they co-operate, she should go to the party and he should follow a mixed strategy assigning a probability of 7/16 to going to the party and 9/16 to staying at home.

A second example illustrates how certain seemingly plausible objections to the principle of minimax relative concession, taken as a basis for co-operation, may be answered. Adelaide and Ernest have the opportunity to co-operate in a mutually profitable way if they can first agree on how to share their gains. Let us suppose that their utilities are linear with monetary values, so that we may give the pay-offs in dollars. Adelaide would receive a net benefit of $500 from their joint venture, provided she receives all of the gains after

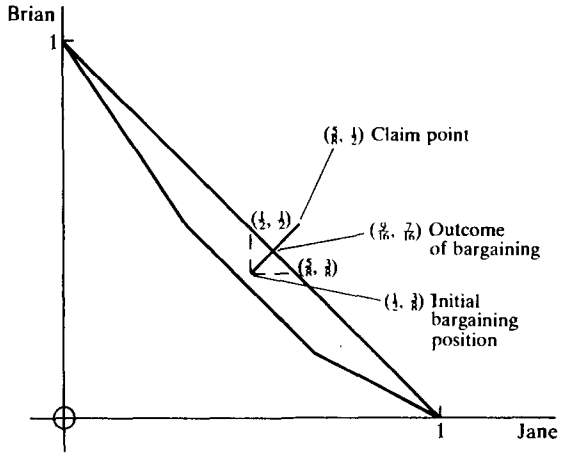

Figure 6

Ernest's costs are fully covered; Ernest however would receive only \$50 if all of the gains go to him after Adelaide's costs are covered. Since their present goods are not in contention, and since they have no alternative sources of gain, we may set the pay-offs to each in the initial bargaining situation at \$0. We suppose that the admissible optimal outcomes fall on the curve shown in Fig. 7.

Each claims as much as affords the other only marginal inducement to co-operate, so that Adelaide's claim approaches \$500 and Ernest's, \$50. From Fig. 7, we see that agreement is possible if each makes a relative concession of nearly 0.3, yielding an outcome affording Adelaide \$353 and Ernest \$35, to the nearest dollar. We see also that agreement on any other outcome would require one to make a greater relative concession. Hence Adelaide should concede \$147 of her possible gain and accept \$353; Ernest should concede \$15 of his possible gain and accept \$35.

Suppose that Ernest complains that Adelaide is getting far more than he—\$353 as opposed to a mere \$35. Adelaide will reply that Ernest is conceding far less than she—a mere \$15 as opposed to \$147. If Ernest were to argue that his gain should be increased because it is so much smaller than Adelaide's, Adelaide would reply that his concession should be increased because it is so much smaller

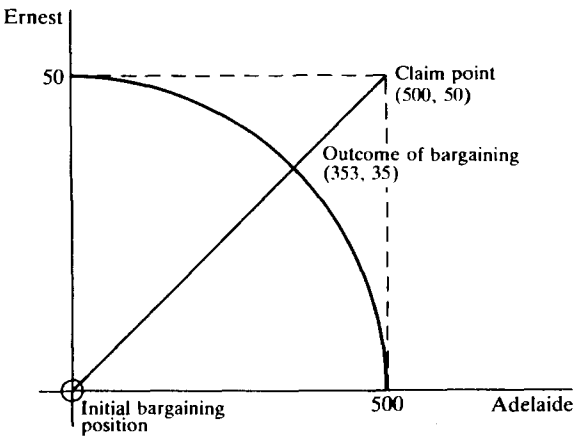

Figure 7

in dollars than hers. If Adelaide were to argue that Ernest should care relatively little about how much he concedes, Ernest would reply that he cares equally little about reaching agreement. His smaller potential benefit reduces the pressure on him to refuse a concession, but equally it reduces the pressure on him to reach agreement and so to make a concession. Her larger stake increases the pressure on her to reach agreement, but equally it increases the pressure on her to hold out against a concession comparable in relative magnitude.

The fundamental rationale for the principle of minimax relative concession does not require even the rough interpersonal comparisons of benefit that we have used in this example. Rather, the rationale turns on an interpersonal comparison of the proportion of each person's potential gain that he must concede. However, were we to assume a measure of utility permitting interpersonal comparisons, and were we then to be tempted by some principle of equal gain, we should remind ourselves that any such temptation could be countered by a principle of equal loss, in relation to one's claim. The unique acceptability of minimax relative concession is made evident when we balance gain in relation to non-co-operation with loss in relation to claim or potential gain.

The two examples we have discussed may suggest that the

principle of minimax relative concession is in fact a principle of minimum equal relative concession—that it requires the smallest equal concessions, measured relatively, from the bargainers. I have indeed claimed this in print, but mistakenly. In most cases, minimax relative concession will result in equal relative concessions, but Alvin Roth has demonstrated that this is not always so.[15]

What we may show is that, if there is an optimal outcome requiring equal relative concessions from each person, then the principle of minimax relative concession selects that outcome. For if an outcome is optimal, then every other outcome in the situation must afford some person a lesser utility. But to accept an outcome with lesser utility, one must make a greater concession—greater in absolute magnitude, since the difference between the utility of one's claim and the utility of one's concession is increased, and so greater in relative magnitude, since the proportion this difference bears to complete concession must also be increased. Thus if there is an optimal outcome requiring equal relative concessions from each person, every other outcome requires a greater concession from some person, and so every other outcome requires a concession greater than the minimax.

There is a simple graphic representation of the outcome required by minimax relative concession, given that there is an optimal outcome requiring equal relative concessions from each. Let the claim point be that point in utility space representing a utility, to each person, equal to what he would expect from his claim. Then all points on the line joining the point representing the initial bargaining position to the claim point, represent outcomes requiring equal relative concession. If this line intersects the upper right bound of the outcome-space, then the point of intersection represents the outcome requiring minimum equal relative concessions, which is the outcome selected by the principle of minimax relative concession. We show this in Figs. 6 and 7.

A third example relates the principle of minimax relative concession to the obvious supposition that in a partnership, each should benefit in proportion to investment. Abel and Mabel agree to pool their resources, finding this advantageous because over the range of their contributions there is a linear increase in returns to scale on investment. Their pooled resources have a value of 1, and they expect a return at rate k. (As in the preceding example, we assume

[15] Roth, pp. 105–7.

utilities are linear with monetary values.) If Abel contributes r, and so Mabel $(1-r)$ to their combined resources, and if Abel would expect a return at rate c were he to invest independently, then by the assumption of a linear increase in returns to scale, Mabel must expect a return at rate $[(k - 2kr + cr)/(1 - r)]$ were she to invest independently. Abel would receive cr from independent investment, and Mabel would receive $(k - 2kr + cr)$. We let these returns be their pay-offs from the initial bargaining position. Abel's claim is then k less Mabel's initial pay-off, or $(2kr - cr)$, and Mabel's claim similarly is $(k - cr)$.

In this situation the principle of minimax relative concession will require equal relative concessions from Abel and Mabel. If we let the return from co-operation to Abel be x, then his absolute concession in his claim $(2kr - cr)$ less x; his full concession is his claim less his initial pay-off; so his relative concession is $[(2kr - cr - x)/(2kr - 2cr)]$. Similarly, Mabel's relative concession is $[(x - cr)/(2kr - 2cr)]$, since her return from co-operation must be $(k - x)$. Since their relative concessions are equal, we may equate them and solve for x. We find $x = kr$. Abel's return kr is equal to the total return k multiplied by the proportion his investment r bears to the total investment 1. And Mabel's return $k(1 - r)$ is equal to the total return multiplied by the proportion her investment $1 - r$ bears to the total investment. These results are of course what we should expect if minimax relative concession does, as we claim, provide the rational basis for co-operation.

3.3 We now present a more formal account of the theory of rational bargaining sketched in the preceding subsections. We begin with some definitions:

(1) The *initial bargaining position* is an outcome, the utilities of which constitute the bargaining endowments of the prospective co-operators. (Thus each bargainer has an entitlement to his utility in the initial bargaining position which is not at stake in the bargaining process.)

(2) The *co-operative surplus* is a set of utility-differences, one for each co-operator, each non-negative in value and equal to the difference between his utility from co-operation and his utility in the initial bargaining position. (The distribution of the co-operative surplus, that is the selection of a particular set of utility-differences constituting such a surplus, is at stake in the bargaining process.)

(3) A *bargaining situation* is a set of outcomes, represented in utility space by a closed, convex figure, the *outcome-space*, and an initial bargaining position, represented by a point of the outcome-space.

(4) A *claim* is a demand by a prospective co-operator for a particular co-operative surplus, made initially in bargaining.

(5) The *claim point* is the point in utility space representing the (possibly hypothetical) outcome that would afford each person a utility equal to that of his claim; it is *feasible* if and only if it is a point of the outcome-space.

(6) A *concession* is an offer by a prospective co-operator to accept a particular utility less than that of his claim.

(7) A *concession point* is a point in utility space representing the (possibly hypothetical) outcome that would result from a set of concessions, one from each prospective co-operator (but possibly including null concessions); it is *feasible* if and only if it is a point of the outcome-space.

(8) A person is willing to *entertain* a concession point if and only if he is willing to make the concession it requires from him provided others are willing to make the concessions it requires from them. The person is then willing to entertain the concession in relation to the concession point.

(9) The *absolute magnitude* of a concession is the difference between the utility of the person's claim and the utility of the concession.

(10) The *relative magnitude* of a concession is the proportion its absolute magnitude bears to the difference between the utility of the person's claim and his utility in the initial bargaining position. Alternatively, the relative magnitude of a concession is the proportion its absolute magnitude bears to the utility-difference for the person of the co-operative surplus that he claimed.

(11) A *maximum concession* for any concession point is the concession (or one of the concessions) with greatest relative magnitude required to bring about the outcome represented by the point.

(12) A *minimax concession* in any bargaining situation is the maximum concession (or one of the maximum concessions) with least relative magnitude required to bring about an outcome represented by a feasible concession point. (Thus in any bargaining situation, each outcome represented by a feasible concession point must require a concession with relative magnitude at least equal to that of the minimax concession.)

We now formulate the conditions on rational bargaining:

(i) *Rational claim.* Each person must claim the co-operative surplus that affords him maximum utility, except that no person may claim a co-operative surplus if he would not be a participant in the interaction required to provide it.

(ii) *Concession point.* Given claims satisfying condition (i), each person must suppose that there is a feasible concession point that every rational person is willing to entertain.

(iii) *Willingness to concede.* Each person must be willing to entertain a concession in relation to a feasible concession point if its relative magnitude is no greater than that of the greatest concession that he supposes some rational person is willing to entertain (in relation to a feasible concession point).

(iv) *Limits of concession.* No person is willing to entertain a concession in relation to a concession point if he is not required to do so by conditions (ii) and (iii).

The rationale for these conditions turns on the benefit each person seeks to realize from the co-operative surplus. Each can increase his utility by co-operating; hence as a utility-maximizer each must find it rational to co-operate. And each recognizes that everyone else must find it rational to co-operate.

Condition (i) is a straightforward application of utility-maximization to the context of bargaining. Each seeks to maximize his return from co-operation so each claims as much as possible, but no person seeks to exclude other co-operators or to be excluded himself, and so each limits his claim to avoid this.

Condition (ii) follows from the fact that, for co-operation to occur, all must agree on an outcome (or joint strategy) represented by a feasible point in outcome-space. Any person who supposes that there is no feasible concession point on which rational persons can agree is denying that there is any way in which those rational persons can co-operate. But each recognizes that everyone finds it rational to co-operate.

Condition (iii) expresses the equal rationality of the bargainers. Since each person, as a utility-maximizer, seeks to minimize his concession, then no one can expect any other rational person to be

willing to make a concession if he would not be willing to make a similar concession.

Finally, condition (iv) is again a straightforward application of utility-maximization, given that the other conditions suffice to require concessions leading to an outcome represented by a feasible point. No rational person can be willing to make unnecessary, or unnecessarily large, concessions. We now demonstrate that conditions (ii) and (iii) do suffice to bring about agreement on a feasible point.

We noted in the parenthetical remark to definition (12) that each outcome represented by a feasible concession point must require a concession with relative magnitude at least equal to the minimax concession. Since by condition (ii) each person must suppose that there is a feasible concession point that every rational person is willing to entertain, then each must suppose that every rational person is willing to entertain a concession point requiring someone to make a concession with relative magnitude at least equal to the minimax concession. But then each must suppose that some rational person is willing himself to entertain a concession with relative magnitude at least equal to the minimax concession, in relation to this concession point. It follows from condition (iii) that each person must himself be willing to entertain a concession with relative magnitude at least equal to the minimax concession, in relation to a feasible concession point.

But in every situation there is a feasible concession point requiring no concession greater in relative magnitude than the minimax concession. By conditions (ii) and (iii), then, every person must be willing to entertain this point. But then condition (ii) cannot require any person to suppose that there is a feasible concession point that every rational person is willing to entertain and that requires a concession greater in relative magnitude than the minimax concession. And so condition (iii) cannot, either itself or on the basis of condition (ii), require any person to be willing himself to entertain a concession greater in magnitude than the minimax concession, in relation to any feasible concession point. Hence by condition (iv) no person is willing to entertain a concession point requiring him to make a concession greater in magnitude than the minimax concession. And so each person must be willing to entertain those and only those feasible concession points that require concessions with

relative magnitudes as great as, but no greater than, the minimax concession.

Let us say that claims that satisfy condition (i) are *maximal*. Then from conditions (i) to (iv) we have established the *Principle of Minimax Relative Concession*: in any co-operative interaction, the rational joint strategy is determined by a bargain among the co-operators in which each advances his maximal claim and then offers a concession no greater in relative magnitude than the minimax concession.

The principle of minimax relative concession plays a threefold role in our argument. First, it expresses the principle of expected utility-maximization in the context of bargaining. Rational bargainers, each seeking to maximize his own utility, determine their claims and concessions by an appeal to the principle. Second, it determines the formal content of a rational bargain; rational bargainers agree on that joint strategy affording each person an expected utility no less than he would expect from his maximal claim and minimax concession. Thus the principle governs both the process and the object of rational choice in bargaining situations.

Third, the principle of minimax relative concession is the principle of rational behaviour in co-operative interaction—interaction based on the joint strategy agreed to in bargaining. Each person acts, not to maximize his own utility, but to bring about the outcome that is the object of the bargain, affording each person an expected utility no less than he would expect from his maximal claim and minimax concession. We have yet to demonstrate that such action is rational—that each person should rationally comply with the joint strategy to which he has rationally agreed. This will be our task in the next chapter. It is this third role which establishes the distinctively moral character of the principle, and of co-operation. For applied to co-operative interaction, the principle of minimax relative concession constitutes a constraint on the direct pursuit of individual utility. Thus if we can show it to be a rational and impartial basis for co-operative interaction, we shall have established its credentials as a moral principle.

Rational persons, faced with the costs of natural or market interaction in the face of externalities, agree to a different, co-operative mode of interaction. They agree to act, not on the basis of individual utility-maximization, but rather on the basis of optimiza-

tion, where the particular optimal outcome is determined by the principle of minimax relative concession. In reaching this agreement, of course, each seeks to maximize his own utility. The same principle, minimax relative concession, serves rational persons both in reaching agreement, and in complying with the agreement reached—both as a principle of choice in bargaining, and as a principle of choice in co-operating. In bargaining minimax relative concession directly expresses the demands of utility-maximization; this we have argued in the present section. In co-operation minimax relative concession constrains and so overrides the demands of utility-maximization. This will be the theme of the next chapter.

3.4 A brief discussion of the principal alternative to our account of rational bargaining, the Zeuthen–Nash–Harsanyi theory, will complete our treatment of this subject.[16] It may be omitted by readers whose interests do not extend to infighting among bargaining theorists, as it contains nothing of positive importance to the development of our argument.

Consider two persons, Ann and Adam, who find themselves in a bargaining situation. Let the initial bargaining position afford Ann a utility u^* and Adam a utility v^*. Ann claims an outcome affording her a utility u_1 and Adam v_1; Adam claims an outcome affording him a utility v_2 and Ann u_2. Assuming their claims to be incompatible, so that (u_1, v_2), the point affording them their claimed utilities, is not feasible, at least one must offer a concession. If Ann were to accept Adam's claim she would give up $(u_1 - u_2)$; if they were to deadlock on the initial bargaining position she would give up $(u_1 - u^*)$. If Adam were to accept Ann's claim he would give up $(v_2 - v_1)$; if they were to deadlock on the initial bargaining position he would give up $(v_2 - v^*)$. The ratio $[(u_1 - u_2)/(u_1 - u^*)]$ measures Ann's loss from accepting Adam's claim as a proportion of her loss from deadlock; the ratio $[(v_2 - v_1)/(v_2 - v^*)]$ similarly measures Adam's loss from accepting Ann's claim as a proportion of his loss from deadlock. We may compare the ratios and suppose, applying Zeuthen's principle, that the person whose ratio is smaller risks more by not making a concession, or alternatively, loses proportionately less by making a concession.

But he need not make a full concession, accepting the other's

[16] See, in addition to the references to Nash and Zeuthen in note 13 above, Luce and Raiffa, pp. 124–37. Harsanyi gives a brief account of earlier work in *Rational Behavior*, pp. 143–53.

claim, if, by making a lesser concession, he may increase his ratio sufficiently so that it is larger than that of the other bargainer. Suppose that Ann must concede; $[(u_1 - u_2)/(u_1 - u^*)]$ is smaller than $[(v_2 - v_1)/(v_2 - v^*)]$. But suppose she can propose an outcome affording her a utility u_3 and Adam v_3, where u_3, although less than her original claim u_1, is greater than what she would receive by accepting Adam's claim, u_2, and also where $[(u_3 - u_2)/(u_3 - u^*)]$ is larger than $[(v_2 - v_3)/(v_2 - v^*)]$. Adam's risk is now greater than Ann's so he must offer a concession.

We suppose that this process of concession continues until their proposals coincide. It can be shown that given this procedure for bargaining, the outcome selected must afford Ann a utility u' and Adam a utility v' such that, for any feasible point (u, v), the product $(u' - u^*)(v' - v^*)$ is at least as great as the product $(u - u^*)(v - v^*)$. Thus on the Zeuthen–Nash–Harsanyi theory, rational agreement maximizes the product of the individual utility-differences in the co-operative surplus.

In some situation this procedure will yield the same outcome as the one we have proposed. But we show in Fig. 8 a bargaining situation where the procedures differ. The possible outcomes are represented by the points (0, 0), (1/2, 1), and (1, 0); (0, 0) is (we suppose) the initial bargaining position. The optimal outcomes are represented by the points on the line joining (1/2, 1) and (1, 0). The

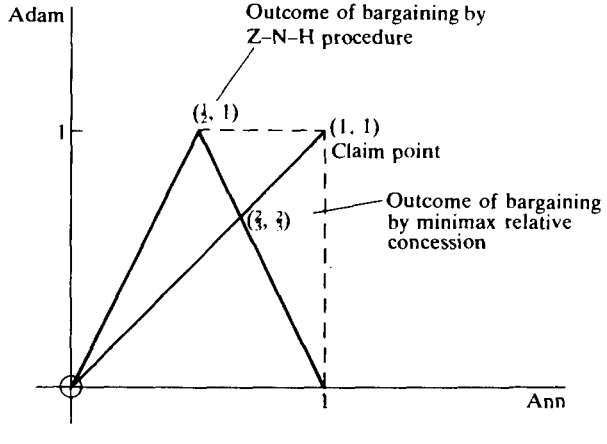

Figure 8

bargaining outcome, by the procedure just sketched, must be represented by that point on the optimal line maximizing the value of $(u - 0)(v - 0)$; this is the point (1/2, 1).

Following our procedure, it is clear that both Ann and Adam must claim a utility of 1; the point (2/3, 2/3) on the optimal line requires each to make a concession with relative magnitude $[(1 - 2/3)/(1 - 0)]$, or 1/3, and any other point, since it affords either Ann or Adam a utility less than 2/3, requires one of them to make a concession with greater relative magnitude. Therefore by the principle of minimax relative concession the bargaining outcome is represented by the point (2/3, 2/3).

We do not suppose that intuition should have the final word in deciding questions in the theory of rational choice. But if it has any word at all, it is surely not spoken here in support of the Zeuthen–Nash–Harsanyi approach, which gives Adam his claim, leaving it to Ann to make all the concessions.

But let us raise deeper objections. The outcome of the Zeuthen–Nash–Harsanyi procedure is determined by maximizing the product of the individual utility-differences in the co-operative surplus. The only constraint that need be placed on the bargainers' claims, to ensure this outcome, is that they specify outcomes that are admissible in the sense of 3.1, and that afford the claimant no less than he would receive from the bargaining outcome. Whereas in our procedure each person has an interest in advancing his maximal claim, on this alternative no one has such an interest, since one cannot expect more by claiming more. There is no rationale for determining the size of one's claim.

At each stage in the process of bargaining a minimum concession is made, in accordance with the requirements of maximizing rationality. But a concession from what? From a claim that has no significance. It is not possible to determine, at the end of the bargaining process, whether one party has conceded more than the other, since concessions are measured only at each stage in the process and are relative to that stage. And even if there be but one stage, we can say that one party has conceded more only in relation to initial claims that, as we have seen, are quite arbitrarily made. Surely then we must question whether the idea of concession retains any real significance on this account of bargaining. In treating the magnitude of claims as arbitrary, the magnitude of concessions also becomes arbitrary.

We might of course suppose that the bargainers were to begin by making real claims—claims satisfying our condition (i). Although they would have no reason to do this, yet if they did, it would enable us to attach meaning to the measure of each person's concessions, by taking these claims as a starting point. But we should then face the obvious and unpalatable consequence that so measured, concessions need not be minimal. Since the outcome of the Zeuthen–Nash–Harsanyi bargaining procedure need not be the outcome yielded by our procedure, we may be certain that it does not always satisfy the principle of minimax relative concession. We may then ask why rational bargainers would employ a procedure that might require greater concessions from one of them than anyone need make.

The core of this alternative theory of bargaining is found in one of Nash's postulates, usually labelled 'independence of irrelevant alternatives'. The postulate, expressed in terms of the representation of the bargaining situation in utility space, states that restricting the outcome-space can affect the bargaining outcome only if it makes the point representing the original outcome unfeasible, outside the restricted outcome-space, or if it affects the initial bargaining position. Alternatively, expanding the outcome-space can affect the bargaining outcome only if the new outcome falls outside the original outcome-space, or again if it affects the initial bargaining position. This postulate makes very explicit the irrelevance of the bargainers' claims, since restricting or expanding the outcome-space may clearly affect the possible claims without affecting the bargaining outcome. Nash's postulate thus denies what we affirm, that the outcome of rational bargaining depends on the claims advanced by the bargainers.

Harsanyi employs a different set of postulates, among which he includes the plausible requirement that the bargaining outcome must not depend (or be expected to depend) on factors whose relevance cannot be established on the basis of the other postulates.[17] In itself this is surely acceptable. But Harsanyi is able to use it to conclude that only those factors that enter into the choice between one's own claim and the other person's claim—between not making and making a concession—are relevant, because of the way in which he describes the bargaining process. If one supposes that fully rational persons proceed by a series of claims (or offers) and concessions leading to new claims, treating each step as involving a

[17] See Harsanyi, *Rational Behavior*, pp. 118, 153–62.

comparative assessment of immediate risks, then unsurprisingly one is led to Harsanyi's conclusion. But if one supposes that rational persons proceed by making maximal claims, and then a single set of concessions, one obtains our result. For what is relevant to bargaining varies between the two descriptions of the process.

One might suppose that the serial character of bargaining, as Harsanyi represents it, is in fact a better account of the true nature of bargaining than we provide. But as we have seen, if one supposes that the serial character of claim–concession–claim is central, then it turns out that from a rational point of view the process is a dummy. What one claims does not matter, and so what one concedes is based on something that does not matter. The outcome of the Zeuthen–Nash–Harsanyi procedure is entirely independent of the serial process that the account purports to capture. On our view, although there is but one claim and one concession on the part of each person, yet the claim does matter, and the concession is then based on something that matters. The process is not a dummy. And this, we insist, more adequately captures the nature of bargaining. We surely suppose that each bargainer's initial claim must be significant, and that his overall concession from that claim must be of central concern. Our account of bargaining satisfies these suppositions. It represents rational persons who are actually bargaining.

4.1 Justice has been silent throughout our long discussion of the internal rationality of co-operation. But as we shall now show, justice and reason coincide in a single ideal of co-operative interaction. The principle of minimax relative concession serves not only as the basis for rational agreement, but also as the ground of an impartial constraint on each person's behaviour. And justice is the disposition to abide by this constraint.

In treating co-operative interaction as the domain of justice we make a twofold claim. The first is that like the market, rational co-operation excludes all partiality. The second is that unlike the market, this exclusion requires each co-operator to constrain her maximizing activity. Interaction that achieves impartiality without constraint constitutes a morally free zone, from which the externalities are absent that lead utility-maximizers into free-ridership and parasitism. But co-operative interaction faces these externalities; co-operation is the visible hand restraining persons from taking advantage of their fellows, but restraining them impartially and in a way

beneficial to all. Such restraint commands rational acceptance; this is the idea underlying morals by agreement.

Our argument in this chapter will be limited in two quite different ways. First, since we have examined only the process of bargaining, leaving aside the initial bargaining position, we shall consider only the impartiality or fairness of this process. Given the initial position, co-operation is just if the joint strategy on which it is based is the outcome of a fair bargain among the co-operators. But the fairness of the bargaining process does not correct any partiality that may be present in the initial position; indeed, it would simply transmit the partiality from the initial position to the joint strategy selected. In Chapter VII we shall examine the conditions under which the inital bargaining position is impartial.

Second, we shall address impartiality, as we have addressed rationality, from the standpoint of the individual actually involved in bargaining. But we noted in Chapter I that other theorists have attempted to link reason and morals by addressing impartiality in an apparently different perspective, considering what principles of interaction would be chosen by any individual occupying a specially conceived, impartial standpoint. Although we shall show in Chapter VIII that this perspective harmonizes with our own, it is not our concern here. We focus on the co-operative choice of a joint strategy, which is impartial because it is acceptable from every standpoint, by every person involved.

Note that no constraint on maximizing behaviour is involved in this choice. Bargaining is a straightforwardly maximizing activity leading to agreement on a joint strategy. Constraint enters in co-operative interaction, which requires adherence to this strategy even though the outcome is not in general an equilibrium. Our concern in this section is to show that if co-operation results from rational agreement, the constraint it imposes is just. But one may still ask, with Hobbes's Foole, whether it is rational to be just, to adhere to the constraint to which one has rationally agreed.[18] This question, of compliance, will be addressed in Chapter VI.

4.2 Recall that co-operation is intended to afford an optimal outcome in situations in which the presence of externalities would make the outcome of natural or of market interaction sub-optimal. The moral significance of externalities is found in the possibility that

[18] See Hobbes, *Leviathan*, ch. 15, pp. 72–3.

one person may take advantage of another, either as a free-rider, obtaining some benefit cost free as a spin-off from the other's activities, or as a parasite, transferring the cost of some benefit to the other. Co-operation, to avoid this possibility, must ensure that the ratio between the benefit the co-operator receives and the contribution she makes is, so far as possible, constant, the same for all. But how is this to be done?

Let us begin with a simple example, in which we shall suppose that the fruits of co-operation—the co-operative surplus, distribution of which is at stake in the bargaining process—may be treated as a single, transferable good.[19] The quantity of this good produced through co-operative interaction is fixed; it may be distributed in any way among the co-operators, and its quantity is not affected. Each person's utilities must be linear with respect to her share of this good, but no interpersonal comparison of utilities is assumed. In our example we shall suppose that this good is money.

Our example is the partnership between Abel and Mabel introduced at the end of 3.2. They expect a return of k on their joint investment. But k is not their co-operative surplus. That surplus is k, less the sum of the return each would expect from independent investment, which for Abel is cr and for Mabel $(k-2kr+cr)$. So the co-operative surplus is $(2kr-2cr)$.

How is this surplus divided between Abel and Mabel? Abel, we showed in 3.2, receives kr from co-operation; subtracting his initial pay-off cr we find that his share of the co-operative surplus is $(kr-cr)$, or exactly one-half. Mabel receives $k(1-r)$ from co-operation; subtracting her initial pay-off $(k-2kr+cr)$ we find that her share of the co-operative surplus is, of course, also $(kr-cr)$. If each receives a total return proportionate to contribution, then each receives half of the co-operative surplus. Since we suppose that returns are divided between Abel and Mabel on the basis of their contributions, this shows that both contribute equally to providing the co-operative surplus.

Our discussion has revealed what must surely appear a very surprising result. If Abel and Mabel are to engage in a partnership in which their returns are proportionate to their contributions, then they must divide the co-operative surplus brought about by their partnership equally. An equal division of the surplus may not seem surprising in itself. For since neither can gain any part of this surplus

[19] For an account of a transferable good, see Luce and Raiffa, pp. 168–9.

without the other, then each is equally responsible for making it available, and so each is entitled to an equal share of it. What is surprising is that this egalitarian view of co-operation proves to be fully compatible with the reasonable insistence that overall returns be proportionate to (differing) overall contributions.

This egalitarian view of co-operation is indeed implicit in our analysis. The co-operative surplus is in the fullest sense the joint product of the co-operators. No one may reasonably or fairly expect more, and no one should reasonably or fairly accept less, than an equal share of the co-operative surplus, where equal shares may be determined. If there is a single transferable good, produced in fixed quantity and divisible in any way among the co-operators, then rationality and impartiality require its equal division. It is evident that this is the outcome required by our theory of bargaining, for given a fixed good, fully divisible, each claims all of it, and maximum concession is then minimized if and only if each person receives an equal share of the good.

But, it may be objected, this result is absurd. Ms Macquarrie, the pharmaceutical chemist, requires a laboratory assistant to aid her in carrying out her experiments, and hires Mr O'Rourke. When as a result of her experiments, Ms Macquarrie discovers a wonder drug that makes her a millionaire, must she divide her royalties with O'Rourke? Of course not. Her experiments were not carried out as a co-operative venture with her assistant. Although she required an assistant, she did not require O'Rourke. Her relationship with him was strictly a market transaction; she hired him at (presumably) the going rate for laboratory assistants. Macquarrie would be ungenerous not to give O'Rourke a handsome bonus, but she does not owe him anything.

Consider, on the other hand, the plight of Sam McGee, the prospector, who discovers the richest vein of gold in the Yukon, but lacks the necessary cash (say $100) to register a claim to it.[20] If Grasp, the banker, is the only man in Dawson City with the ready cash to lend to McGee, then poor Sam will (rationally) have to offer Grasp a half-share in the claim. For although Grasp's $100 is worth only $100 in the absence of McGee's discovery, yet McGee's discovery is worthless to him without Grasp's money. Of course, if there are other sources of funds, Sam is in a position comparable to

[20] For S. McGee, see 'The Cremation of Sam McGee', in R. W. Service, *Songs of a Sourdough* (Toronto, 1907).

Ms Macquarrie; he needs money but not Grasp's money, so he can borrow at the going rate. But in the Dawson City of our example there is, alas, no going rate.

4.3 What if there is no single, transferable good, produced in fixed quantity and divisible at will among the co-operators? How may we determine fair shares of the co-operative surplus? Recall the significance of the claim each person makes. Each claims as much of the surplus as it is possible for him to receive, but only for those co-operative interactions in which he participates. Thus no contribution yields no claim; some contribution yields full claim. A full claim is a claim to the entire surplus in so far as the claimant can receive it. Let us then equate the full claims of the bargainers; each makes a claim equivalent to that of every other person. Suppose that all receive the same proportion of their claims. Then since their claims are equivalent, they receive equivalent shares of the co-operative surplus.

We suggested in the preceding subsection that each person who contributes to a co-operative interaction is equally responsible for the resulting surplus. We may modify this, if co-operation does not result in a fixed quantity of a fully transferable good, to say that each person has a responsibility equivalent to what he can receive of the co-operative surplus. Each then has an equal entitlement to the surplus in so far as the products of co-operation are available to him. Although we have no measure enabling us to divide the surplus into equal shares, yet we can divide it (in general) into shares which are equivalent in terms of the full claims, and so entitlements, of the bargainers. Thus if a fair or impartial distribution of the co-operative surplus relates the benefit each person receives to the contribution he makes, each person's fair share of the surplus is determined by making shares proportional to claims.

We are now in a position to set out a principle of fair bargaining. Suppose, as in 3.2, that the initial bargaining position affords some person a utility u^*, and he claims an outcome affording him a utility u^*. Then his claim for the co-operative surplus is $(u^* - u^*)$. Suppose the joint strategy chosen for co-operative interaction affords him an expected utility u; then his expected share of the co-operative surplus is $(u - u^*)$. The proportion of his claim to the co-operative surplus that he receives is then $[(u-u^*)/(u^*-u^*)]$; we shall call this his *relative benefit*. Note that relative benefit must fall between 0, which is the

relative benefit of the initial bargaining position $[(u^*-u^*)/u^*-u^*)]$, and 1, which is the relative benefit of the claim $[(u^*-u^*)/(u^*-u^*)]$. If we are to make shares proportional to claims, then we must divide the co-operative surplus so that each person receives maximum equal relative benefit.

This may not always be possible. Or rather, it may be that if each person receives the maximum equal relative benefit, then not all of the co-operative surplus will be distributed among the co-operators. In such a case, it would seem reasonable to maximize minimum relative benefit. Since this is equivalent to maximizing equal relative benefit when the latter uses up the co-operative surplus, we may say that a fair or impartial distribution of the co-operative surplus—and so a fair or impartial choice of the joint strategy for co-operative interaction—must accord with the *Principle of Maximin* (maximum-minimum) *Relative Benefit*.

But this is our old friend, minimax relative concession, in a new guise. For relative benefit and relative concession sum to unity. $[(u-u^*)/(u^*-u^*)] + [(u^*-u)/(u^*-u^*)] = 1$. One maximizes minimum relative benefit by minimizing maximum relative concession. Impartiality and rationality coincide in bargaining.

We should not find this result surprising. For those considerations that might justifiably lead us to question the fairness of many instances of everyday bargaining cannot arise in the context of full rationality. In ordinary bargaining persons may conceal significant features of their circumstances, or the full range of their options, may misrepresent their preferences, or the strengths of their preferences. But we suppose each person to be fully informed—to know the possible actions available to every person, the possible outcomes that may result from those actions, and the utility pay-offs to every person of each possible outcome. In ordinary bargaining persons may bluff, especially if they are also able to conceal or misrepresent factors, so that others have uncertain or mistaken expectations about what the bluffers are willing to do. But here there is no place for bluffing; not only is each person fully informed but he is a rational utility-maximizer who knows his fellows to be also rational utility-maximizers. In ordinary bargaining persons may make threats, but among fully rational persons threats are useless; no one will believe anyone who claims that he will act in a non-utility-maximizing way should others not comply with his threat, and to

say that one will act in a utility-maximizing way is not to threaten.[21] Our bargainers have no psychological strengths to exploit, or psychological weaknesses to be exploited. And we assume that bargaining is cost free, in terms of both utility and time, so that no one need come to a decision without full consideration; bargaining is unpressured. Thus each bargainer can employ only his own rationality to appeal to the equal rationality of his fellows. In addition to rationality, there are only each person's preferences and possible actions to consider, and it is about these that everyone bargains.

We might suppose that even granting the ideal nature of our bargainers, their circumstances might nevertheless give rise to significant partiality. But problems about circumstances concern the initial bargaining position, not the process of bargaining. We shall turn to these in Chapter VII. Certainly, if the ideal of co-operative interaction is to eliminate the free-ridership and parasitism of natural interaction, then some constraint on identifying the initial position with the non-co-operative outcome will be required, since that outcome may incorporate the effects of the very activities co-operation would eliminate. But here our concern is with the choice of a basis for co-operative interaction taking an initial position for granted, and our argument is that in affording equal or equivalent shares of the co-operative surplus to all persons, the principle of minimax relative concession ensures that bargaining impartially relates each person's contribution to co-operation to the benefit he receives from it.

Let us conclude this discussion by noting that many of our actual moral principles and practices are in effect applications of the requirements of minimax relative concession to particular contexts. We may suppose that promise-keeping, truth-telling, fair dealing, are to be defended by showing that adherence to them permits persons to co-operate in ways that may be expected to equalize, at least roughly, the relative benefits afforded by interaction. These are among the core practices of the morality that we may commend to each individual by showing that it commands his rational agreement.

[21] We shall argue in VI.3.2 that threats are compatible with rationality, but not in the context of co-operation. In effect, threat behaviour would be proscribed by the constraint on the initial bargaining position, to be developed in Chapter VII.

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 début 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,

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

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

[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.

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 any one, 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

[8] Ibid., ch. 14, p. 65.
[9] Ibid.
[10] Ibid., ch. 21, p. 111.
[11] Ibid., ch. 15, p. 71.
[12] Ibid.

> 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 errour of them that receive him; nor when he is received, be retayned in it, without seeing the

[13] Ibid., ch. 15, p. 72.
[14] Ibid., ch. 15, p. 73.

danger of their errour; which errours 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

[15] Ibid.

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 Foole'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 Foole.

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.

[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.

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 non-compliance utility. The benefits that could be realized through co-operative 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

[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.

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 Foole 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 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.

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

[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

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-co-operation. 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 co-operation in relation to mutual non-co-operation, but each benefits from non-co-operation 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-co-operative outcome.

Constrained maximizers can thus obtain co-operative benefits that are unavailable to straightforward maximizers, however far-sighted 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

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 Co-operation among Egoists', *American Political Science Review* 75 (1981), pp. 306–18.

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 co-operative 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-co-operatively, 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 manœuvre. Since our argument is to be applied to ideally rational persons, we may simply add another idealizing assumption, and take our persons to be *transpar-*

ent.[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

[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.

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, co-operation, 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-co-operation 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 co-operation 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-co-operative 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-co-operation, 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-co-operation to exploitation. And so, since defection has the value 1 and exploitation 0, let us assign co-operation the value 2/3 and non-co-operation 1/3.

The gain from defection, $(1-u')$, thus is 2/3; the gain through co-operation, $(u''-u')$, 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 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. 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-co-operation, 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, $(u'' - u')$, 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'', decreases, and so the degree of translucency required to make co-operation 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-co-operation, *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 co-operation 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 clear-headed 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 co-operate, 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 Fooles, it would not pay to be a constrained maximizer, and to comply with one's

[21] Hume, *Enquiry*, iii. i, p. 187.

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 Fooles 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 Foole 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

[22] Ibid., ix. ii, pp. 282–3. [23] Ibid., ix. ii, p. 283.

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

[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.

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 satisfice, 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

[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.

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.[27] 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

[27] See R. L. Trivers, 'The Evolution of Reciprocal Altruism', *Quarterly Review of Biology* 46 (1971), pp. 35–57.

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 co-operative 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 co-operation 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 2.1—constrained maximization is not straightforward maximization in its most effective guise. The

[28] Ibid., p. 50.

[29] J. Elster, *Ulysses and the Sirens: Studies in rationality and irrationality* (Cambridge, 1979), p. 145.

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.

[30] Ibid., pp. 145–6.

VII

THE INITIAL BARGAINING POSITION: RIGHTS AND THE PROVISO

1 Once upon a time, long ago and far away, there was a society of masters and slaves. Unlike many such societies, this one rested on no false ideological appeals to the natural masterliness of masters and the natural slavishness of slaves. What distinguished the masters from the slaves, as both well knew, was power. Given half a chance, the slaves would happily have changed places with their masters, who therefore were careful never to give them that half chance.

But one day, as a group of masters were taking their customary leisurely mid-afternoon gin and tonics together, one of the younger men, recently returned from the university, interrupted the desultory remarks of his elders and held forth thus:

'Gentlemen, we have been fools these many years. The meagre pleasures that have consoled us [and here he snapped his fingers that his glass might be refilled] have been but shadows of the luxuries we might have enjoyed, had we not squandered our resources in coercing our slaves. And those slaves have performed their tasks grudgingly and carelessly (dammit, boy, where's the *lime* in my drink?) under the threat of our whips and chains, instead of freely and cheerfully serving us. To use the language of my professors, we've interacted non-co-operatively and ended up with a sub-optimal outcome.

'Now we can change all this. What we need is a bargain with our slaves—we'll free them, dismantling all the coercive apparatus that slavery involves, and in return they will voluntarily be our servants. We'll benefit—better service, less expense, money saved for other uses, and they'll benefit—no more beatings and chainings, and better living conditions, since from the money we'll save in doing away with coercion and the increased productivity we'll get from their willing service, we can pay them wages better than the living allowance we have to provide now, and still have the resources to put some *real* pleasures into our lives. I've figured it all out—there's this professor who's developed what he calls the principle of minimax relative concession for rational bargains. This deal with our slaves ought to fit it just about perfectly. And he's shown—this professor—that it's rational to comply voluntarily with such deals. So we shouldn't have any worries about doing away with coercion—our slaves are rational so they'll become willing servants.'

The older men shook their heads. Not only did a university education cost a lot of money, but it left one with the damnedest fool ideas—anyone with half an eye would see through a scheme like that. Take away the chains and you'd never get willing servants. But the younger men, impressed with the prospect of rational co-operation, were not to be put off by old fogies who thought that grandfather's way of doing things was best. At the very next election (for the masters were quite democratic among themselves) a reform administration carried the day, and within a few years the institutions of slavery were dismantled, an emancipation proclamation issued, and a solemn Bargain of Mutual Benefit enshrined in the constitution.

And did it all happen as the young man had said? Not at all. The older men had been quite right. As one of the ex-slaves explained, after he was sworn in as prime minister of an administration pledged to repeal the Bargain of Mutual Benefit, 'What those young men never understood was that this bargain was coercively based. It was only because of the power they held over us that it seemed a rational deal. Once that power was taken away, it became obvious that the fruits of co-operation weren't being divided up in accordance with that fancy principle of minimax relative concession. And so there wasn't any reason to expect voluntary compliance—we weren't about to become willing servants. Still, they saved themselves a revolution—so in the end it probably was a sensible move for them to make, even if they did it because of their mistaken expectations.'

This tale sets the scene for our discussion of the initial bargaining position—or indeed more generally, of the initial position whether for co-operative or market interaction. In Chapters IV and V we saw that both market competition and agreed co-operation begin from a specification of the factor endowments and prior utility expectations of the parties to interaction. We established the procedural rationality and fairness of these forms of interaction. But we noted explicitly that fair procedures yield an impartial outcome only from an impartial initial position. And it is equally true that rational procedures yield a rationally acceptable outcome only from a

rationally acceptable initial position. Implicit in the prime minister's remarks in our cautionary tale is the claim that it is rational to comply with a bargain, and so rational to act co-operatively, only if its initial position is non-coercive.

In section 2 we shall develop this claim, examining two views that permit coercion in the initial position, and arguing that they do not provide a basis for rational compliance. The first view identifies the non-co-operative outcome of natural interaction with the initial bargaining position; the second view identifies the threat point—the outcome yielded by each person's maximally effective threat strategy—with the initial position. We shall deny the relevance of threat behaviour to rational interaction, and argue that, if the non-co-operative outcome involves coercion, then it must be constrained by removing the effects of that coercion if it is to serve as an initial position for bargaining to a joint strategy that rationally commands individual compliance.

But removing the effects of coercion from the non-co-operative outcome does not afford an adequate positive characterization of the initial position. The principal claim that we shall advance and defend in the present chapter is that whether we assess the rational acceptability of the outcome of interaction from the standpoint of each individual utility-maximizer, or the moral acceptability of the outcome as benefiting each individual impartially, we find that all effects of taking advantage must be removed from the initial position. We shall therefore argue that it is both rational and just for each individual to accept a certain constraint on natural interaction, and on the determination of his initial factor endowment, as a condition of being voluntarily acceptable to his fellows as a party to co-operative and market arrangements—to social interaction. This constraint is part of morals by agreement, not in being the object of an agreement among rational individuals, but in being a precondition to such agreement.

In developing this argument, we shall proceed from a characterization of this constraint to the demonstration of its moral and rational grounds. We begin then in section 3 with the statement of a form of the Lockean proviso, that there be 'enough, and as good left ... for others'.[1] We show how the proviso functions as a constraint,

[1] The term 'Lockean proviso' comes from R. Nozick, *Anarchy, State, and Utopia* (New York, 1974), who discusses it on pp. 175–82. The proviso is stated in John Locke, *Two Treatises of Government* (London, 1690), second treatise, ch. v, paras. 27, 33.

but we do not show in section 3 that it actually determines initial factor endowments, that it is impartial, or that it is rational to adhere to it. In section 4 we turn to the first two of these tasks, showing how adherence to the proviso introduces a structure of rights into a previously non-moral state of nature, and arguing that this structure satisfies the core moral standard of impartiality. And then in section 5 we show that it is rational for utility-maximizers to accept the proviso as constraining their natural interaction and their individual endowments, in so far as they anticipate beneficial social interaction with their fellows.

This last point is essential. We may say that the proviso moralizes and rationalizes the state of nature—but only in so far as we conceive the state of nature as giving way to society. Although it is irrational for human beings to remain in a state of nature, accepting no constraints on their interactions, yet no individual can benefit from unilaterally constraining his behaviour. Adherence to the proviso is the equivalent of the requirement in Hobbes's first law of nature, 'That every man, ought to endeavour Peace, as farre as he has hope of obtaining it'.[2] But without such hope, of passing from nature to society, then every man 'may seek, and use, all helps, and advantages of Warre'. Without the prospect of agreement and society, there would be no morality, and the proviso would have no rationale. Fortunately, the prospect of society is realized for us; our concern is then to understand the rationale of the morality that sustains it.

2.1 The identification of the initial bargaining position with the non-co-operative outcome has been defended by James M. Buchanan in his convincing outline of a contractarian theory of rational interaction.[3] Since we agree with much of Buchanan's approach we must consider his argument carefully.

Buchanan supposes for simplicity a two-person world with one scarce good. This good is not produced by the inhabitants; it 'falls down' on them in fixed quantities. They cannot benefit from trade with each other, or from investment in production. Nevertheless we may suppose some interaction between them, dictated by the desire each has to consume as much of the good as possible. Each may have, or believe that she has, reason to invest effort in obtaining some portion of the good that originally fell down on the other. In

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

[3] See J. M. Buchanan, *The Limits of Liberty: Between Anarchy and Leviathan* (Chicago, 1975).

other words one or both may invest in predation. And this investment may give one or both reason to make a counter-investment in defence. The eventual result of this predatory/defensive interaction is the emergence of what Buchanan calls the natural distribution, a condition of stability in which the predatory/defensive mix of each individual is the utility-maximizing response to the other's mix. This natural distribution is, in our terminology, the non-co-operative outcome.[4]

But now there is a basis for agreement. The natural distribution 'serves to establish an identification, a definition, of the individual persons *from which* contractual agreements become possible. Absent such a starting point, there is simply no way of initiating meaningful contracts, actually or conceptually.'[5] We may say that the natural distribution affords each person an explicit bargaining endowment; it determines what each may bring to the table and thus constitutes an initial bargaining position. Of course, the mere existence of a natural distribution is not sufficient for contracts to emerge. It must be sub-optimal—there must be the possibility of mutual improvement. But its sub-optimality is hardly in doubt, since the effort expended in predation and defence is largely wasteful. Both parties stand to benefit from an agreement that relieves them of the necessity of engaging in these non-productive activities. (If it be pointed out that under Buchanan's simplifying assumptions, there is no place for production, we may reply that there is place for leisure.)

Let us then suppose that agreement is reached, proceeding from the natural distribution, and let us also suppose that, in accordance with our argument in Chapter V, this agreement is based on the principle of minimax relative concession. (The search for a principle of agreement is no part of Buchanan's concern.) Then each person will receive her share in the natural distribution, or its equivalent in utility, plus a proportion of her potential benefit from co-operation, in general equal to that of the other person. Is this rationally acceptable to each party?

The answer is surely negative. In Buchanan's example, the sub-optimality of the natural distribution results from investment in predation and counter-investment in defence. Each imposes costs on the other by these activities. Although pure predation, in the form of unhindered seizure, need not be wasteful, yet the cost imposed on the predator by the defensive response her predation elicits, and the

[4] See ibid., pp. 23–5.

[5] Ibid., p. 24.

cost of that defensive response, are both unproductive. Agreement ends this unproductive activity. It yields an optimal outcome in which predatory/defensive efforts are absent. But the effect of these efforts remain present, since each party brings to the bargaining table the fruits of her predatory/defensive activity, and takes them (or their utility equivalents) away from the table as part of the overall outcome. They do not enter into the co-operative surplus, which is constituted (in Buchanan's example) entirely by the agreement of each party to cease imposing costs on the other. Only this comes under the sway of the principle of minimax relative concession.

But clearly an individual would be irrational if she were to dispose herself to comply, voluntarily, with an agreement reached in this way. Someone disposed to comply with agreements that left untouched the fruits of predation would simply invite others to engage in predatory and coercive activities as a prelude to bargaining. She would permit the successful predators to reap where they had ceased to sow, to continue to profit from the effects of natural predation after entering into agreements freeing them from the need to invest further predatory effort. Co-operative compliance is not compliant victimization. We do not deny that, as long as her cost in resisting actual predation exceeds any benefit such resistance would bring her, the victim rationally must acquiesce. But if predatory activity is banned, then she no longer has reason to behave in a way that would maintain its effects. Agreement reached by minimax relative concession from the natural distribution therefore does not elicit the rational, voluntary compliance of both (or all) parties, if the natural distribution is in part the result of coercion.

Our initial tale is intended to illustrate this argument. The masters employ coercion to keep the slaves obedient. Coercion is costly to both. Masters and slaves would both benefit were coercion removed and the slaves continued to serve voluntarily. But ex-slaves would not comply with an agreement to this effect. The slaves provide their services because the costs of their resistance exceed the benefits it would afford them, given their masters' power. But only the maintenance of this power rationally induces them to continue their services. Without coercion, ex-slaves might accept and adhere to some form of co-operation, but not one based on the outcome of coercive interaction. If it were otherwise, then why should not an agreement, of the type proposed by the young master, be con-

cluded—and adhered to—by whites and blacks in South Africa, or by the government of Poland and the members of Solidarity?

Buchanan is not unmindful of this problem. He asks 'Why will persons voluntarily comply with the rules and institutions of order that are in being?', where voluntary compliance excludes that based on coercion and punishment.[6] He insists, 'This question can only be answered through an evaluation of the existing structure, *as if* it were the outcome of a current contract, or one that is continuously negotiated. Individuals must ask themselves how their own positions compare with those that they might have expected to secure in a renegotiated contractual settlement.' And these positions are determined, at least in part, by considering 'imagined shifts in the natural distribution in anarchistic equilibrium which always exists "underneath" the observed social realities'.[7]

Buchanan supposes that one should comply with those rules (or, we might say, with that joint strategy) that would command agreement were the present underlying natural distribution to be realized. In deciding whether to comply one asks, 'Were agreement to lapse, then what might I expect?' Buchanan depends on the threat implicit in the natural distribution to elicit compliance. But a return to the natural distribution benefits no one. The threat is unreal. What motivates compliance is the absence of coercion rather than the fear of its renewal.

Consider once again the simple two-person world Buchanan discusses, and which we represent graphically in Fig. 9 (see p. 228). An agreement from the point I_n, the natural distribution or non-co-operative outcome, would lead to the optimal point B_n. But B_n requires individual V to make an unproductive transfer to person U, corresponding to U's net predatory gain from natural interaction. V would have to give U, quite voluntarily, some of the good that, in Buchanan's example, falls down on him. We have argued that in the absence of coercion V has no reason to do this. If he refuses, keeping the good for himself, then the effect is to move the outcome along the frontier, or upper right bound of the outcome-space, to B_c, the optimal point representing the outcome of agreement from the non-coercive initial position I_c. If U objects, seeking to compel a transfer from V, then she must reintroduce coercion, and so the outcome moves along the line joining B_c to the natural distribution I_n, a move which is costly to both parties and so irrational for U. Hence we

[6] Ibid., p. 75.

[7] Ibid., p. 79.

must suppose that a stable agreement ends at B_c even if the parties actually begin at I_n. The eventual outcome is as if I_c were the initial bargaining position.

In Buchanan's simple world, each will simply enjoy the good falling down on her, and devote to leisure the effort previously invested in predation and defence. But in a more complex world, we may suppose that this effort may be put to productive use, so that the two parties enter either into market arrangements or into co-operative production in which they obtain goods previously unavailable. The expected gains from their agreement are then considerably increased. And so if V refuses to make an unproductive transfer and U seeks to enforce it, then U must not only reintroduce coercion, but also give up the goods achieved from the productive use of the efforts that agreement had enabled her to divert from predation and defence. The cost to U, and so the irrationality of seeking to compel an unproductive transfer from V, becomes even more evident.

To relate this elementary analysis to our society of masters and slaves, we note that the point I_n corresponds to the situation at the beginning of our tale. B_n represents the outcome of agreement as naïvely imagined by the young master, in which without coercion the ex-slaves continue to serve their ex-masters. Provision of this service is the unproductive transfer required by this agreement from the ex-slaves, which of course they refuse. Their refusal moves society from B_n to B_c, the eventual outcome after the repeal of the so-called Bargain of Mutual Benefit. And this outcome is as if agreement had been reached from a non-co-operative but also non-coercive initial position, represented by the point I_c.

We have argued that it would be irrational for an individual to dispose herself to comply voluntarily with a joint strategy agreed to from a coercive initial position. In the light of our analysis we may reformulate our claim to say that it would be irrational for an individual to dispose herself voluntarily to make unproductive transfers to others. An unproductive transfer brings no new goods into being and involves no exchange of existing goods; it simply redistributes some existing goods from one person to another. Thus it involves a utility cost for which no benefit is received, and a utility gain for which no service is provided.

It is rational to make an unproductive transfer only if it is directly utility-maximizing to do so. Since the transfer itself is costly, it can be utility-maximizing only in so far as it is coercively exacted. It

cannot, then, be part of any co-operative interaction, since such interaction involves mutual benefit. The presence of unproductive transfers in otherwise co-operative arrangements is evidence of residual natural predation.

In our examples residual predation may easily be detected. But in our world it readily disguises itself. Unproductive transfers parade in spurious moral and ideological trappings. If we are no longer taken in by the blandishments of nobility, we are all too ready to succumb to the plaints of inequality. But ideally rational persons are not so moved, recognizing unproductive transfers for what they are. They may coerce and be coerced, but they do not confuse coercion with co-operation.

Since the outcome of the co-operative joint strategy is not in equilibrium it lacks natural stability. Compliance provides an artificial but weaker stability where the stronger natural stability is not to be had. If we were to take the natural distribution as the initial bargaining position, then in some cases we should find that unproductive transfers were necessary to bring about the co-operative outcome. Even if particular transfers could be coercively exacted within co-operative arrangements, yet the arrangements themselves ultimately depend on voluntary compliance. And this would not be forthcoming from those called on to make the transfers. The co-operative outcome would then lack both natural and artificial stability. Hobbes's Foole would be able to show that, even though it might seem rational to enter into co-operative interaction, it would not be rational to comply with its demands.

We conclude that the initial bargaining position, as the starting point for rational co-operation, may not be identified with the natural distribution, or non-co-operative outcome. The natural distribution represents the effects of power. Now we have not shown, and we shall not show, any basis, rational or moral, for criticizing the effects of power considered in themselves. But if we consider the natural distribution in relation to market or co-operative interaction, then we assess it, not in itself, but for its suitability as determining what each person brings to the market or to the bargain underlying co-operative arrangements. We assess it as determining each person's endowment. And here we have found a basis for criticizing it, and indeed rejecting it in so far as it is coercive. If we think of each person's endowment as constituting her rights (a view we shall develop in 4.1 of this chapter), then we may

say that to accept the natural distribution as the initial bargaining position would be to accept might as making right.[8] No doubt right is often invoked to maintain what might has made. But the right so invoked is an impostor, unable to pass the scrutiny of utility-maximizing rationality.

2.2 In the theory of rational bargaining developed by John Nash and generalized by John Harsanyi, the initial bargaining position is identified with the *threat point*, representing the outcome that would be realized were each person to act on her maximally effective threat strategy.[9] What are these strategies?

We saw in V.3.4. that on the Zeuthen–Nash–Harsanyi view, if the initial bargaining position for two persons, Ann and Adam, is represented by the point (u^*, v^*), then the outcome must be represented by that point (u', v') such that for any point (u, v) in the outcome-space, the product $(u'-u^*)(v'-v^*)$ is at least as great as the product $(u-u^*)(v-v^*)$. In this way we may correlate each point in the outcome-space, considered as an initial bargaining position, with a point on the upper right bound of the outcome-space, as the bargaining outcome for that initial position.

Conversely, we may correlate each point on the upper right bound of the outcome-space with that set of points in the outcome space each member of which (taken as initial bargaining position) determines it as the bargaining outcome. It is clear that each point in the outcome-space belongs to one and only one such set; two points belonging to the same set determine the same outcome; two points belonging to different sets determine different outcomes. Let us call these *outcome-equivalent sets*.

It is clear that in choosing a point to serve as the initial bargaining position, Ann and Adam are both indifferent between all points belonging to the same outcome-equivalent set, but have strictly opposed preferences between any two points belonging to different sets, corresponding to their strictly opposed preferences between any two optimal outcomes. In choosing a point to serve as the initial bargaining position, Ann and Adam are concerned with, and only with, the outcome-equivalent set to which it belongs.

[8] Buchanan insists that the initial bargaining position 'cannot properly be classified as a structure of *rights*, since no formal agreement is made' (ibid., p. 24). But we think of rights as providing the basis rather than the object of agreement; see 4.3 *infra*.

[9] See J. F. Nash, 'Two-person Cooperative Games', *Econometrica*, 21 (1953); Luce and Raiffa, pp. 140–3; J. C. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge, 1977), pp. 167–9.

Suppose that, although the strategies available to Ann and Adam remain unchanged, the pay-offs of the outcomes are changed, so that what we shall call the utility^c of each outcome is a measure of preference for the outcome-equivalent set to which it belongs. (Preference for each set is of course determined by preference for its associated bargaining outcome.) Then it turns out that Ann and Adam face a strictly competitive situation, in which each has a strategy, utility^c-maximizing against the other's strategy, and so leading to an outcome in equilibrium^c, which is also optimal^c. These are their maximally effective threat strategies; if they were acting to determine an initial bargaining position, these are the strategies they should rationally choose.

This very neat analysis cannot be applied directly to our solution to the bargaining problem, but fortunately we need not face the problem of determining maximally effective threat strategies. For consider what these strategies signify. They play a purely hypothetical role in the Nash–Harsanyi analysis, since Ann and Adam do not actually choose them, but merely appeal to them to determine the costs that each could impose on the other in a strict competition for bargaining advantage. Maximally effective threat strategies would not be chosen by Ann and Adam were they to find themselves unable to co-operate; the threat point bears no particular relationship to the non-co-operative outcome. But if Ann and Adam would not choose these strategies, then they cannot credibly threaten with them. Maximally effective threat strategies prove to be idle.

Bargaining theorists generally suppose that individuals are in a position to make their threats binding.[10] But this is an unrealistic supposition in most situations. And we have already seen that in co-operation, threat behaviour would be proscribed. Thus, even apart from the irrationality of complying voluntarily with an agreement that took the threat point as the initial bargaining position, we conclude that potential co-operators would not dispose themselves to threat enforcement. We may dismiss the threat point from further consideration.

3.1 The initial bargaining position must be non-coercive. But must we go further in constraining natural interaction, in so far as it determines the basis of market or co-operative interaction? We shall argue that the terms of fully rational co-operation include the

[10] Harsanyi, however, distinguishes situations with and without binding threats. See *Rational Behavior*, p. 168.

requirement that each individual's endowment, affording him a base utility not included in the co-operative surplus, must be considered to have been initially acquired by him without taking advantage of any other person—or, more precisely, of any other co-operator. Otherwise those who consider themselves taken advantage of in initial acquisition will perceive society as unfair, in demanding payments from them without offering a compensating return, and will lack sufficient reason to accept market arrangements or to comply voluntarily with co-operative joint strategies.

Our concern in this and most of the fourth section will be to understand the requirement. We shall consider what is meant by the taking of advantage, and show how not taking advantage constrains the acquisition of an endowment for market and co-operative interaction. In our discussion we shall seek to make plausible the claim that not taking advantage is a reasonable and fair constraint that natural interaction must satisfy in so far as its outcome provides an initial position for bargaining. Thus we shall understand not taking advantage to exclude free-riders and parasites, so that as a constraint it extends to the state of nature the impartiality satisfied by the market and by co-operation. But plausibility is not proof; after we examine the effects of not taking advantage of others in natural interaction we shall seek to demonstrate that it is a moral and rational requirement. And we must emphasize again that nothing in our discussion should be taken to imply that it would be rational to refrain from taking advantage of others, or to adhere to any other constraint on straightforward maximization, from the strict standpoint of interaction in a state of nature. Throughout, we envisage not taking advantage as constraining natural interaction only in order to facilitate the emergence of society.

The first great attempt to rationalize and moralize the Hobbesian state of nature—that condition of unlimited predation—is John Locke's theory of property.[11] The principle of acquisition that constitutes the core of this theory suggests a preliminary formulation of the requirement that advantage not be taken. But we shall extend Locke's own position so that we may express his constraint on acquisition as a proviso that simultaneously licenses and limits the exclusive rights of individuals to objects and powers. Its effect is to afford each person a sphere of exclusive control by forbidding others from interfering with certain of his activities. This exclusive

[11] See Locke, second treatise, ch. v, and first treatise, ch. ix.

sphere constitutes a moral space, which defines the individual in his market and co-operative relationships.

According to Locke one acquires exclusive title to that with which one mixes one's labour, provided one uses, or at least does not waste, what one so acquires, and provided also that 'enough, and as good' is left, 'more than the yet unprovided could use'.[12] The initial bargaining position determined by this principle thus involves a distribution of factors of production and other goods based on each person's natural labour, in so far as this labour is related to use, and in so far as each person's acquisition leaves a surplus. Labour, use, and surplus are the three key concepts in understanding Locke's account, and each may seem deeply problematic.

Locke's theory of acquisition moralizes the Hobbesian state of nature. But how is this possible? How does my labour, to which I have only a liberty according to Hobbes, serve as the basis of my rights? How is Locke able to show that I have a right to my body and its powers, except by presupposing it and so begging the question of establishing a rational and impartial bargaining position? To resolve these difficulties we must approach the idea of initial or original acquisition in terms of the surplus constraint, which we shall find to be its core. The acquisition of one's body and its powers, the role of labour, and the demand that one put to use what one acquires, will all fall into place, given a suitable reading of the Lockean proviso expressed in the words 'enough, and as good', which constrains natural interactions in order to make society possible.

Taken literally these words might seem to defeat our enterprise from the outset. If one must leave enough and as good for others, then may one claim anything at all for one's factor endowment? Even the acquisition of one's body may seem to be precluded, for if you are abler and stronger than I, then your claim to your body and its powers will not leave enough and as good for me. I should do better with a half-share in our joint capacities than with a full share in mine and none in yours. And beyond this, to require enough and as good to be left would surely forbid the acquisition not only of what is actually scarce, but also of what is merely potentially in insufficient supply. There cannot be enough of what is scarce, so that if one person takes sufficient for himself, what is left for others is not as good. This literalistic reading of Locke's proviso would simply

[12] Locke, second treatise, ch. v, para. 33.

fail to define persons for the purposes of bargaining—or the market. The world and all of its inhabitants would be and remain a commons.

We are therefore led to consider the interpretation of the proviso offered by Robert Nozick. 'Locke's proviso that there be "enough and as good left in common for others" . . . is meant to ensure that the situation of others is not worsened.'[13] Natural interaction, if it is to determine the initial bargaining position, must exclude activities that worsen the situation of any person, whether by predation or in other ways. Each person's endowment includes whatever he acquires without worsening the situation of his fellows; this endowment then affords him a basis for applying minimax relative concession and so determines the expected outcome of co-operation. Each is free to use the resources in his endowment to increase what he brings to the bargaining table, and also, of course, to engage in market competition.

But simply to forbid worsening the situation of others is too strong. For there are situations in which one could avoid this only by worsening one's own position. Following Locke who allows one's own preservation to take justifiable precedence over that of others in one's deliberations, we modify Nozick's interpretation of the proviso, so that it prohibits worsening the situation of others except where this is necessary to avoid worsening one's own position. It is clear that any stronger proviso would fail as a possible constraint for determining the initial bargaining position among utility-maximizers.

How are we to understand worsening—and, conversely, bettering—someone's situation, where that someone may be oneself or another? We may treat 'better' and 'worse' as unproblematic; one situation is better for some person than another, if and only if it affords him a greater expected utility. Now one's situation is bettered or worsened only in relation to some base point. This base point cannot be that of the initial bargaining position itself, since that position is to be determined by the application of the proviso forbidding worsening the situation of another except to avoid worsening one's own. We must ask what, in natural interaction, may be taken as a base point for determining what effects the actions of one person have on the situation of another, or on his own situation.

[13] Nozick, p. 175.

The crucial distinction that we must establish is between worsening someone's situation and failing to better it, since the proviso prohibits only the former, not the latter.

To aid our enquiry at this point, let us consider an example beloved of philosophers. You are drowning in the river, and I, passing by on the river bank, leave you to drown. This is an outcome; consider two ways in which it might have come about. First, you fall into the water. I come along, hear your cries for help, but ignore them and continue on my way. Second, you are standing on the bank. I come along, push you into the water, and, ignoring your cries for help, continue on my way. In the first case, although certainly I fail to better your unhappy situation, I do not worsen it. In the second case, although the outcome is the same, I clearly do worsen your situation. Why this difference?

Suppose that I had not come along. Then in the first case you would likely have drowned; in the second case you would have remained safely on the bank. The outcome in the first case is no worse for you than it would have been in my absence; hence I have not worsened your situation. However, on the assumption that I might have been able to save you, or to help save you, from drowning, the outcome in the first case might have been better for you than it would have been in my absence, had I acted differently; hence I have failed to better your situation. (Were I, for whatever reason, completely incapable of doing anything to help you, then I should not have bettered your situation, but I should not have failed to better it.) The outcome in the second case is worse for you than it would have been in my absence; hence I have worsened your situation.

In this simple example, the base point for determining how I affect you, in terms of bettering or worsening your situation, is determined by the outcome that you would expect in my absence. Worsening, and equally bettering, are judged by comparing what I actually do with what would have occurred, *ceteris paribus*, in my absence. Failing to better, and equally failing to worsen, are judged by comparing what I might have done, but did not do, with what would have occurred without me. I push you in; you would not have fallen in without me; I worsen your situation. I could have saved you; you would not have been saved without me; I fail to better your situation.

There are complications to the determination of the appropriate base point which this example does not capture. If you are drowning, and I am a life-guard on duty, then by ignoring your cries for help I do worsen your situation. For here my behaviour is to be judged against what would have happened in my absence on a normally life-guarded beach, not on a beach that in lacking me also lacked a life-guard. But note that failure to provide a life-guard, assuming no prior requirement or commitment, could not worsen or contribute to worsening the situation of users of the beach, even though, once a life-guard is provided, her inattention to duty may worsen their situation. Complications such as this indicate the importance of considering the institutional framework and the nature of practices within which actions occur, in order to determine the base point appropriate for judging bettering or worsening.

To this point we have considered only the effect of one person on the situation of another. But what if a person affects her own situation? Here we cannot appeal to her absence in order to fix the base point. A determinate context for bettering or worsening one's own situation is however provided by the assumption of interaction. Although we may speak of someone bettering her situation in so far as she prefers its outcome both to her previous position and to some alternative that she might have brought about, yet this is not relevant to our present discussion. Rather, just as I better your situation in so far as you prefer the outcome of interaction with me to what you would have expected otherwise, in my absence, so I better my own in so far as I prefer the outcome of interaction with you to what I should have expected otherwise, in your absence or unavailability for interaction. And I worsen my own situation in interaction with you in so far as I prefer what I should have expected in your absence to the outcome that I actually bring about. Again, this abstracts from any institutional framework.

We interpret the Lockean proviso so that it prohibits worsening the situation of another person, except to avoid worsening one's own through interaction with that person. Or, we may conveniently say, the proviso prohibits bettering one's situation through interaction that worsens the situation of another. This, we claim, expresses the underlying idea of not taking advantage.

3.2 The proviso is intended to apply to interaction under the assumptions of individual utility-maximizing rationality and mutual unconcern. Each person is supposed to choose a strategy that maximizes his expected utility, unless specifically forbidden by the

proviso to do so. Each is then free to better his own situation as he chooses, provided that he does not thereby worsen the situation of another. To require that, as a condition of bettering one's own situation, one must better that of others, would be to require that one give free rides. But no one is free to better his own situation through interaction worsening the situation of another. To allow that, in order to better one's own situation, one may worsen that of others, would be to allow one to be a parasite. Thus a stronger constraint on natural interaction than the proviso would license free-ridership, a weaker constraint would license parasitism. Navigating between these, the proviso constrains natural interaction to make rational, fair, and free co-operation possible.

A person who accepts the proviso as a constraint on his maximizing behaviour may be supposed to reason about the choice of a strategy in interaction in the following way:

(1) He divides his strategies into three groups:
A: those that afford each person an expected utility no less than she could expect in the absence of interaction;
B: those that afford him an expected utility no less than he could expect in the absence of interaction, but afford some other person an expected utility less than she could expect in the absence of interaction;
C: those that afford him an expected utility less than he could expect in the absence of interaction.

(2) If Group A is not empty, then he chooses from it a strategy maximizing his expected utility. He betters, or at least does not worsen his situation, without worsening that of anyone else.

(3) If Group A is empty and Group B is not empty, then he chooses from Group B a strategy minimizing the loss to others of expected utility in comparison to what they could expect in the absence of interaction. He does not worsen his own situation, and compatibly with this he minimizes worsening the situation of others. The proviso forbids unnecessary worsening.

(4) If Groups A and B are empty, then he chooses from Group C a strategy maximizing his expected utility. He minimizes worsening his situation when interaction is unprofitable to him.

Consider an example of interaction constrained by the proviso. Joanna and Jonathan find themselves castaways on a small and

otherwise uninhabited island. Concerned to establish a basis for co-operation, each resolves to comply with the proviso. Thus Jonathan does not force Joanna to submit to his sexual desires, even though it might better his situation (in relation to her absence), because it would worsen hers (in relation to his absence). But when Joanna expresses her willingness, Jonathan maximizes his gratification without concern for her arousal. He betters himself without worsening her situation. Later, of course, they may agree to enhance each other's sexual enjoyment, finding this mutually beneficial. But such mutual enhancement is co-operative.

Each takes whatever food he or she pleases, without concern for the other. Since in the other's absence each would have access to all the food the island provides, each would worsen his or her situation by refraining from taking whatever food he or she wants and can get. However, when Joanna begins to cultivate a garden and grows a greater supply of fruits and vegetables, Jonathan refrains from seizing what he pleases, since he would then be bettering himself (for in her absence there would be no garden) by worsening her situation (since in his absence she would enjoy the garden alone). And when Jonathan makes a fishing rod and catches a greater supply of seafood, Joanna, for like reason, refrains from seizing what she pleases.

Of course, by this time we should expect Joanna and Jonathan to have reached an agreement to co-operate—and a good thing, too, given the pronounced swelling in Joanna's abdomen. But only such an agreement—explicit or implicit—can bring them to an optimal state of affairs. In natural interaction, even constrained by the proviso, each assesses every situation separately in deciding what to do, so that only immediate reciprocity can be taken into account. In co-operation, they may extend their horizon of concern, accepting costs now for expected benefits later.

In natural interaction it would be irrational for Joanna to run any risk whatsoever to save Jonathan from the sharks in the lagoon—assuming, of course, mutual unconcern. Were she to worsen her situation by bettering his in this way, she would in effect allow him to be a parasite. But an understanding to co-operate may place each of them under a rational and moral obligation to endeavour to save the other even at real risk to oneself. For both Joanna and Jonathan stand to benefit from a practice imposing certain costs on each in order to confer greater benefits on the other. Each insures himself or

herself against being left to whatever perils there may be by agreeing to seek to rescue the other.

The proviso, then, does not incorporate the demands of fair optimality into the state of nature. It merely constrains natural interaction to determine an initial position from which a fair and optimal outcome may be attained. Or so we claim—for remember that we have yet to show that the proviso satisfies the standards of morals and reason.

4.1 Locke introduces the proviso to constrain a particular activity, the acquisition or appropriation of external objects. He argues that the exercise of this activity in conformity with the proviso affords the actor an exclusive right to the land or goods appropriated. In this way the actor obtains the material endowment in the factors of production that is required fully to define him for the purposes of market and co-operative interaction. But Locke assumes that each person begins with an exclusive right to his body and its powers, a right which is extended through labour to what he acquires.

For us the proviso plays a wider and more basic role. We treat it as a general constraint, by which we may move from a Hobbesian state of nature, in which there are no exclusive rights whatsoever but only liberties, to the initial position for social interaction. We begin by considering what each person may do with that body and those powers to which he has direct access, and what others may and may not do with that body and those powers. We show that each person has certain rights to his body and powers, constituting his basic endowment as we defined it in IV.3.2. We then consider how these rights, which Locke assumes but which we derive, lead to further rights in land and other goods. These arise, not simply from acts of appropriation conformable to the proviso, but jointly from what a person may do and from what others in consequence may not do—from the way in which an individual may use certain goods and the consequent restrictions that preclude others both from using the goods and from interfering with his use.

In this way we show that the proviso introduces a rudimentary structure of rights into natural interaction. It converts the predatory natural condition described by Hobbes into the productive natural condition supposed by Locke. But its primary role is to make possible the further structures required for the forms of social interaction, both competitive and co-operative. Of course we could

imagine other ways of making social interaction possible, other ways of establishing a structure of rights determining each person's initial endowment. Our claim is that the proviso accomplishes this in a way both rational and impartial.

In the Hobbesian state of nature one may use one's body and powers as one pleases. Although the proviso constrains this use, it does so without affecting the core of the liberty manifested. For in exercising one's powers one need not interact with others, and so one need neither better one's situation in relation to what one would expect in their absence, nor worsen their situation in relation to what they would expect were one absent oneself. Any constraint arising from the proviso affects the manner in which one exercises one's powers, but not the mere fact of that exercise.

In the Hobbesian state of nature one may also use the bodies and powers of others as one pleases. Within the bounds of what one is able to do there is nothing one may not do. But this liberty is directly affected by the proviso. For in using the powers of others one betters, or expects to better, one's situation through interaction with them, and in so far as one's use must interfere with their own exercise of their powers, one worsens their situation by that interaction.

Each person, in the absence of his fellows, may expect to use his own powers but not theirs. This difference is crucial. For it provides the base point against which the proviso may be applied to interaction. Continued use of one's own powers in the presence of others does not in itself better one's situation; use of their powers does better one's own situation. Refraining from the use of one's own powers worsens one's situation; refraining from the use of others' powers fails to better one's situation but does not worsen it. Continued use of one's own powers may fail to better the situation of others but does not in itself worsen their situation; use of others' powers, in interfering with their own use, does worsen their situation. Thus the proviso, in prohibiting each from bettering his situation by worsening that of others, but otherwise leaving each free to do as he pleases, not only confirms each in the use of his own powers, but in denying to others the use of those powers, affords to each the exclusive use of his own. The proviso thus converts the unlimited liberties of Hobbesian nature into exclusive rights and duties. Each person has an exclusive right to the exercise of his own powers without hindrance from others, and a duty to refrain from

the use of others' powers in so far as this would hinder their exercise by those with direct access to them.

We have provided the justification, promised in IV.3.2, of the *basic endowment*. We suppose that each person identifies with those capacities, physical and mental, to which he has direct access, and we see that this identification affords each person a normative sense of self, expressed in his right to those capacities. By appealing to the proviso we show that this identification is not arbitrary, but rather fully justified by the (yet to be shown) rationality and impartiality of the proviso.

The first step in the conversion of the state of nature is now complete; application of the proviso affords each person exclusive right to the use of his body and its powers, his physical and mental capacities. The next step is to extend this right to the effects of exercising one's powers. Here Locke's concern with labour and use enter the argument. But as we shall see, this second step does not afford an exclusive right to the fruits of one's labour, even if one puts those fruits to use.

Suppose that in the state of nature I cultivate a plot of land, intending to consume its produce. Here my exercise of my powers is quite independent of any other person, and so I do not better myself through interaction. Even if I worsen the situation of someone who would otherwise have cultivated the land, this worsening is incidental to the benefit I receive. My activity cannot violate the proviso. Now suppose that some other person seizes the produce of the land which I have cultivated. Then she does better herself as a result of my activity, and furthermore worsens my situation from what it would have been in her absence, by depriving me of the fruits of my labour. Her activity does violate the proviso. Thus we see more clearly what we claimed in the example of 3.2; Jonathan must not seize Joanna's fruits and vegetables, and Joanna must not seize Jonathan's fish.

Note that to establish a violation of the proviso, we appeal both to my labour in producing the good that the other person takes, and my intended use of that good. She betters her position through my labour, and worsens my position by depriving me of the expected use. The proviso would not be violated were she to take a windfall that I had thought to use but not laboured to produce, nor would it be violated were she to take produce for which I could find no use whatsoever.

By this argument we demonstrate a right in the effects of one's labour, but not an exclusive right to their possession. For we have not shown that the proviso would be violated were someone to seize the fruits of my labour while compensating me for my effort and intended use. If the benefit I receive is no less, in terms of my utilities, than what I expected from my labour in the absence of intervention, then my situation has not been worsened.

Thus the proviso requires what, following Nozick, we term *full compensation*, and not *market compensation*.[14] Full compensation leaves a person without any net loss in utility. Market compensation affords her a share of the benefit realized by the other individual who seizes the good and pays compensation—the share that could be expected from voluntary exchange. One worsens the situation of another in not giving her full compensation for the effects of one's actions on her. One would better the situation of another in giving her market compensation (where this exceeds full compensation); the gain from exchange is a benefit she could receive only through interaction. The proviso prohibits worsening but does not require bettering another's position, in bettering one's own. It does not then require that the person who seizes the fruits of another's labour share her gains; it requires only that she compensate for costs. Thus the proviso affords a right *in* the fruits of one's labour and so to full compensation, not a right *to* those fruits and so to market compensation.

This second step, from a right to one's body and its powers to a right in one's products, completes the conversion of the pure state of nature. The remaining steps concern the transition from natural interaction to market and co-operative interaction. The first of these constitutes the internalization of costs necessary to the emergence of the market.

Suppose that we live as fisherfolk along the banks of a river. If you compel me to fish for you then you violate the proviso in preventing me from exercising my powers as I see fit. If you seize the fish I catch then you violate the proviso unless you compensate me for my labour and my intended use of the fish. But if you, living upstream from me, merely use the river for the disposal of your wastes, then even though you thereby kill many of the fish in my part of the stream, you do not violate the proviso. For although you worsen my situation in relation to what I should expect in your absence, you do

[14] See ibid., pp. 57, 63–5.

not better your own situation through interaction with me. You are no better off than you would be were no one to live downstream from you. The cost you impose on me is not necessary to the benefit you receive; it is not a *displaced cost*. Rather, the cost is occasioned solely by my presence, which from your point of view may be simply unwanted.

But suppose now that we cease to live as independent fisherfolk. Instead of consuming all of the fish you catch, you use some in trades with or involving me. My willingness to trade—my desire for fish and the terms on which I accept fish—are of course affected by the supply of fish directly available to me, and so by your polluting activity. Exchanging with me betters your situation; from your point of view interaction with me is profitable. But it may not better my situation, taking as base point your absence. Although I benefit by trading with you when the alternative is not trading, yet I may do less well than I would were I alone, fishing in an unpolluted stream. Taking all of the ways in which we interrelate into account, you better your situation through interaction that worsens mine. And so your use of the river for waste disposal, because of its effect on the terms of trade between us, violates the proviso.

Suppose however that taking our interaction as a whole, each of us improves his situation. Our exchanges would then seem to compensate me—fully—for the pollution you cause. Does this free you from the charge of violating the proviso? No; the absence of global worsening does not show that no part of our interaction violates the proviso. You dispose of your wastes in a way that kills the fish in my part of the river. You thereby impose a cost on me that betters the terms of trade for you and correspondingly worsens them for me. The cost you impose on me is now necessary to some part of the benefit you receive, and so it is a displaced cost. You benefit from polluting my water; you better your situation through interaction that worsens mine.

Our exchanges do not constitute compensation to me for your pollution. You do not trade with me in order to compensate me. You do not trade with me because you have polluted my water; your method of waste disposal is not a necessary condition of our being able to trade. Indeed, our exchanges, far from compensating me for your pollution, are less beneficial to me than they would have been, were it not for your pollution. If you talk of compensation here, you add insult to injury. You impose uncompensated costs on me by

your method of waste disposal, and given our interdependence, you benefit from imposing those costs and so violate the proviso.

The proviso is violated by an action that betters the actor's situation through worsening the situation of another person. An action performed in one context may be entirely innocent; it may benefit the actor and impose costs on another person, but benefit and cost are quite unrelated. The same action performed in another context may be clearly guilty; it may benefit the actor and impose costs on another person, and the benefit, at least in part, may depend on the cost.

By polluting the water in the stream in which I fish, you increase my demand for your fish, and so improve, from your standpoint, the terms of trade between us. But this need not be the only benefit you gain at my expense. Suppose that we are no longer simple fisherfolk, but two industrialists. We need pure water in our manufacturing activities, and we produce waste which is most cheaply disposed of by discharging it into water, thereby rendering the water impure. If each of us produces only for his own use, your use of the stream for waste diposal, although increasing my costs of production (since I must now install water-purifying equipment in my plant), does not decrease yours in relation to what they would be in my absence. But suppose each of us produces to sell to a third party. Now you put yourself at an advantage by keeping your waste disposal costs down in a way that raises my production costs. In this market situation, any costs of your activity that fall on me are displaced costs, benefiting you by worsening my competitive position, and so, if uncompensated, constitute violations on the proviso.

We may generalize from this argument to conclude that in both market and co-operative interaction, all of the costs of one person's activities that fall on others within the sphere of interaction are displaced costs, requiring compensation if the proviso is not to be violated. For even if, narrowly conceived, some of these costs may seem unnecessary to the particular benefits received, yet in so far as the framework of the market or of co-operative institutions and practices is required for those activities, all costs occurring within that framework are necessary. To be sure, no displacement of costs can occur within the perfectly competitive market, or within co-operative interaction governed by the principle of minimax relative concession, but their initial positions are the outcome of natural interaction that must satisfy the proviso if no one is to take initial

advantage. Thus in so far as natural interaction leads to market or co-operative interaction, the full internalization of costs among those interacting is required by the proviso. Even if in the state of nature itself, certain costs arise only because of the presence of others who do not affect the benefit received, yet in so far as these others enter into social interaction, costs imposed on them in the state of nature must be compensated in determining their market and co-operative endowments.

Because of the importance of this point to our analysis, we reiterate it. In a pure state of nature, the imposition of costs on those incidental to one's activity can in no way violate the proviso. Such costs are strictly incidental and unnecessary to the benefits one receives. One does not then better one's situation through interaction, even though one does worsen the situation of those on whom the costs fall. But if one views those persons as potential partners in social relationships, in market competition and in co-operation, then the proviso forbids the imposition of any costs upon them without appropriate compensation. For now the costs falling on them put them at a disadvantage with respect to the envisaged relationships, and so betters one's situation by worsening theirs.

Full, not market, compensation is required. Although our concern is with preconditions of the market, it is not with market interaction. The fisherfolk have a right in their stream, but not—yet—a right to it. Such rights emerge only in the final step by which the state of nature is converted into society. The effect of the first three steps in rationalizing and moralizing the state of nature is the creation of a framework of common use among interacting persons, in which full compensation is required should one person interfere with the use another makes or proposes to make of certain material goods, but in which there are no exclusive rights of possession to external objects. The fourth, and final step, which defines the full endowment of each individual, introduces exclusive rights to land and other goods.

Suppose that several persons inhabit an island. The land and its resources comprise in effect a commons available to all. But use is individual; each person provides primarily for her own needs, and interaction is non-co-operative. To make our account more realistic we should think not of individual persons but of families. The idea of a family is of a group the members of which take an interest one in another; hence internal interaction within the family is not treated

directly in our analysis. We suppose then that each family provides for its own needs, interacting non-co-operatively with other families. But one individual, or head of a family, aware that planned, intensive cultivation would make the land more productive, proposes to take a certain area of the island for her exclusive use, so that she (and her family) may benefit by maximizing its productivity. She seeks an exclusive right to a certain portion of the island.

How may we assess this proposed right? First we must ask whether someone, in seeking exclusive use of land or other goods, violates the proviso, bettering her situation through worsening that of others. If not, then we must ask whether some other person, in interfering with a claim to exclusive use, violates the proviso. If so, then the proposed right is established. The right-holder, and no one else, may use the land or goods without violating the proviso. And if the proviso is violated, the compensation required will be, not full, but market compensation, for the right-holder is entitled to as much as she could have obtained by voluntary exchange, as compensation for any use made of her land or goods. A right *to* land or goods is a right not only to the fruits of use, but also to the fruits of exchange.

We begin then by considering the effects of granting a claim to exclusive control. On the one hand it is evident that the person, whom we shall call Eve, intends to better her situation by her appropriation of land for herself. And she intends to better it in relation to the base point set by the terms of the problem—that is, in relation to the system of common use. She seeks the security of tenure that a right to market compensation, rather than merely full compensation, confers. Thus she intends to better her situation in relation to her fellows. And this conclusion is reinforced by noting that a system of rights determines the endowments for prospective market and co-operative interaction; exclusive right to a portion of land therefore affords Eve a more favourable basis for such interaction than she would otherwise enjoy.

But although Eve intends to better her situation in relation to her fellows, she need not seek to bring this about by worsening their situation. They are, it is true, to lose their right to use in common the land that she appropriates. They are to be obliged to enter into exchanges with her that she voluntarily accepts, rather than merely paying her full compensation, should they use what she produces. Now we might suppose that Eve seeks a part of the island so large that she would leave her fellows worse off than before; the land

remaining in common might support them less well than the entire island supported everyone. Eve's claim would then violate the proviso. But she need not seek such a large appropriation. Planned intensive cultivation made possible by her security of tenure may well make it possible for her to live better on a part of the island sufficiently small that the others would also be better off, living without her on the remaining land, than they were when all used the entire island in common. For of course, in seeking a private holding, Eve proposes to give up her right in the remaining commons.

Furthermore, the other inhabitants of the island may also benefit from new opportunities to trade their products for some of the goods resulting from Eve's more intensive cultivation. She may produce sufficient food to meet the needs of several families, so that others, who formerly grew their own food, may become specialist craftsmen and craftswomen, with benefits to all. Hence her appropriation may enable everyone to improve her situation, in relation to the base point set by use in common, so that it does not violate the proviso.

Given that the effects of exclusive use would be mutually beneficial, we may now consider the effects of interference with Eve's claimed right. This interference must take one of two forms; it may tend to restore common use, or it may involve only a transfer of exclusive use. Given the benefits of Eve's appropriation, the restoration of common use could not but worsen the situation of most persons. The benefits brought about by Eve's security of tenure would no longer be forthcoming. Any benefits that the person seeking to restore common use might hope to obtain would be purchased at the expense of most of his fellows, in clear violation of the proviso. The transfer of exclusive use—the seizure of the land from the original appropriator—is even more evidently in violation. The person seizing the right is bettering himself by worsening Eve's situation, and may avoid this only by paying market compensation, negotiating with Eve for the right to the land on mutually acceptable terms. But then Eve's right is recognized. If she does not violate the proviso by her claim, then anyone subsequently interfering with that claim would violate the proviso.

Eve's right is thus vindicated. Exclusive rights of possession may afford benefits to all, because they give individuals the security needed for it to be profitable to themselves to use the resources available to human beings in more efficient and productive ways.

They transform a system in which each labours on a commons to meet her own needs into a system in which each labours on her own property and everyone's needs are met through market exchange. Individual self-sufficiency gives way to role specialization. The division of labour opens up new ways of life, with opportunities and satisfactions previously unimagined. Thus the mutually beneficial nature of exclusive rights of possession provides a sufficient basis for their emergence from the condition of common use which is the final form of the state of nature. These rights depend on the proviso, which allows individual appropriation and forbids subsequent interference. Each person stands fully individuated and defined in relation to her fellows, her right to her own body and powers now extending to a right to material goods. Where the market fails, co-operative modes of interaction are available to achieve optimality in the face of externalities. Eve, the first appropriator of property, is the great benefactress of humankind, although in this she is led by an invisible hand to promote an end which was no part of her intention.

Different persons will of course benefit differentially from the emergence of a system of exclusive rights. We may assume that Eve, who first takes land for her exclusive use, will take the best portion; no other person is then able to make an equally advantageous appropriation. Eve does not leave her fellows 'as good' to appropriate, although in taking for herself she leaves them as well off, and indeed better off, than before. The proviso ensures that at every stage in interaction, each person is left as much as she could expect from the previous stage. Advantage is thus not taken, but equality is not assured. We must show, then, that the inequality allowed by the proviso is no indication of partiality.

4.2 Is the proviso an impartial constraint on interaction? What might lead us to question its impartiality? Does it go too far, in prohibiting persons from bettering themselves through interaction that worsens the situation of others? Does it not go far enough, in leaving persons free to better themselves as long as they do not gain from worsening the situation of others? Or in leaving persons free to worsen the situation of others, if the only alternative would be to worsen their own? Or does the proviso fail to be impartial, because it is based on considerations about bettering and worsening? It is this last question which raises the serious issues we must address. It is the relevance of bettering and worsening which is challenged, in questioning the impartiality of the proviso.

For, it will be urged, the proviso says nothing about equalizing. Or, it will also be urged, the proviso says nothing about meeting needs. The rich man may feast on caviar and champagne, while the poor woman starves at his gate. And she may not even take the crumbs from his table, if that would deprive him of his pleasure in feeding them to his birds.

Distressing as we may find this situation, we should not be misled by it. We think of rich and poor within a social context, and we think that his wealth and her poverty are in some way related. If so, then in examining how the situation came about, we may well find a violation, if not of the proviso, then of the principle of minimax relative concession. We should begin with a quite different example—one that we shall adapt to our purposes from Robert Nozick's *Anarchy, State, and Utopia*.[15]

There are sixteen Robinson Crusoes, each living on a different island. Each is either clever or stupid, either strong or weak, either energetic or lazy. Each lives on an island either well or ill supplied to meet human needs and desires. No two Robinson Crusoes are alike in their characteristics and circumstances (so given sixteen of them, all possible combinations are realized). There is a clever, strong, energetic Crusoe living very comfortably on a well-supplied island. There is a stupid, weak, lazy Crusoe barely surviving on an ill-supplied island. In between are fourteen other Crusoes.

Each is equipped with a two-way radio, putting him in communication with the other Robinson Crusoes. Each knows how his situation compares with that of each of the others. Each is also able to build small rafts, not large enough to carry a man or woman, but sufficient to carry provisions of various kinds. The ocean currents will take these rafts from one island to another, but only in a single direction, so that trade is impossible.

Suppose now that the clever, strong, energetic Crusoe, on the well-supplied island, could send provisions by raft to the stupid, weak, lazy Crusoe, on the ill-supplied island. It would no doubt be generous of her to do so. She might want to do so; there is nothing in our story to require that she takes no interest in the other Crusoes' interests. But suppose she does not want to do so; she knows about the others, but does not care about them. Is she under an obligation to supply the stupid, weak, lazy Crusoe? Is it unfair of her not to do so? Is the principle that allows each Crusoe to use his own capacities

[15] See ibid., p. 185.

and the resources of the island on which he finds himself for his own benefit, an unfair principle, or one which expresses partiality to some persons? Would an impartial principle require the better situated Crusoes to contribute to the worse situated ones, or the able Crusoes to the less able ones, where the ocean currents made such contributions possible? Would it require equalizing contributions when possible? Or would it set out needs and require that these be met when possible? No; any principle other than the one allowing each Crusoe to benefit himself would be unfair and partial, in requiring some to give free rides to others, or to be hosts for their parasitism.

Time passes; the Crusoes learn how to make larger rafts, making migration (in the direction of the currents) a possibility. If we follow the proviso, each has a right to his physical and mental capacities but only a right in the resources of the island he inhabits. (Given the difference between well-supplied and ill-supplied islands, once migration is possible some would better themselves by an exclusive right to an island in a way that would worsen the situation of others.) Hence the Crusoes may migrate, and each may use the resources of any island on which he lands, although he must compensate (fully) any other Crusoe if he takes the fruits of the other's labour. The clever, strong, energetic Crusoe on the ill-supplied island may (currents permitting) move to the well-supplied island occupied by a stupid, weak, lazy Crusoe and, ignoring the latter, make himself as comfortable as his counterpart who began her life on a well-supplied island. Of course some migrations may lead to co-operation—the clever but weak and lazy Crusoe supplying the brain for the strong and energetic but stupid Crusoe's brawn, for example. But they need not. Is the principle which allows each Crusoe to use his own capacities and the resources of any island he can reach for his own benefit an unfair one? Would an impartial principle require abler Crusoes to move (currents permitting) in order to assist less able ones? No; such a requirement would be unfair, as again requiring giving free rides or hosting parasites.

John Rawls insists, 'There is no more reason to permit the distribution of income and wealth to be settled by the distribution of natural assets than by historical and social fortune.'[16] We should rather 'regard the distribution of natural talents as a common asset and . . . share in the benefits of this distribution whatever it turns out to be. Those who have been favored by nature, whoever they are,

[16] J. Rawls, *A Theory of Justice*, (Cambridge, Mass., 1971), p. 74.

may gain from their good fortune only on terms that improve the situation of those who have lost out. . . . No one deserves his greater natural capacity nor merits a more favorable starting place in society.'[17] It is clear that Rawls would require persons to bargain from a position of equality, not only with respect to initial rights to goods, but also with respect to initial rights to personal powers and capacities. Those who are more capable or more fortunate than their fellows must not only refrain from taking advantage of others, as the proviso requires, but they must give advantage to others as a condition of benefiting themselves. For Rawls morality demands the giving of free rides; no other interpretation can be put on the insistence that talents be treated as a common asset.

We may agree with Rawls that no one deserves her natural capacities. Being the person one is, is not a matter of desert. But what follows from this? One's natural capacities determine what one gets, given one's circumstances, in a condition of solitude. One's natural capacities are what one brings to society, to market and co-operative interaction. Why should they not determine, or contribute to determining, what one gets in society? How could a principle determine impartially how persons are to benefit in interaction, except by taking into account how they would or could benefit apart from their interaction?

Rawls talks of the 'distribution of natural talents' and of the 'natural lottery'.[18] He falls prey to the dangers that lurk in this talk. There is no natural lottery; our talents are not meted out to us from a pool fixed to guarantee winners and losers. And if there is a distribution, there is no distributor—unless we assume a theistic base foreign to Rawls's argument. The proviso determines the initial endowments of interacting persons, taking into account the real differences among those persons as actors. But the proviso is not itself an actor or a distributor; it enters into the agency of persons in so far as they respect it. If there were a distributor of natural assets, or if the distribution of factor endowments resulted from a social choice, then we might reasonably suppose that in so far as possible shares should be equal, and that a larger than equal share could be justified only as a necessary means to everyone's benefit. But this would be to view persons as the creatures of a distributor—a God or a non-instrumental Society—and not as rational and individual

[17] Ibid., pp. 101–2.

[18] Ibid., pp. 101, 74.

actors. In agreeing with Rawls that society is a co-operative venture for mutual advantage, we must disagree with his view that natural talents are to be considered a common asset. The two views offer antithetical conceptions of both the individual human being and society.

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

If sixteen Robinson Crusoes lived, each on a separate island, and if each used his capacities to provide for himself from the resources of the island, then the outcome, whatever it might be, could not be unjustified. No more can it be unjustified if each person brings her capacities, and what she has realized through them without worsening the situation of others, as her individual endowment for the competitive and co-operative endeavours that constitute society. The proviso, in determining the rights persons have on the basis of what they do, and in treating what persons do from the standpoint of the individual actor, ensures the impartiality of interaction.

4.3 Human beings have rights. Moral theory has not been notably successful in providing an account of these rights. Utilitarianism, as we saw in IV.4.1, leads inexorably to the view that 'rights . . . are . . . vested in the government and their association with individuals is a temporary matter of convenience'.[19] If this view were not unacceptable on other grounds, we should have to ask how governments come to acquire or possess rights, a question not easily answered if one supposes governments to be instruments rather than masters of their citizens, but we may draw a decent veil of obscurity over this embarrassment to utilitarians. Forthright defenders of individual rights have on occasion been content to put the task of grounding rights to one side, 'following the respectable tradition of Locke, who does not provide anything remotely resembling a satisfactory explanation of the status and basis of the law of

[19] Winch, p. 99.

nature'.[20] Locke, indeed, unlike his modern-day disciples, did not suppose that a secular grounding of individual rights was possible; his moral theory, unlike Hobbes's, is overtly theistic.

Contractarianism offers a secular understanding of rights. But the idea of morals by agreement may mislead, if it is supposed that rights must be the product or outcome of agreement. Were we to adopt this account, we should suppose that rights were determined by the principle of minimax relative concession. But as we have seen, the application of this principle, or more generally, the emergence of either co-operative or market interaction, demands an initial definition of the actors in terms of their factor endowments, and we have identified individual rights with these endowments. Rights provide the starting point for, and not the outcome of, agreement. They are what each person brings to the bargaining table, not what she takes from it.

Market and co-operative practices presuppose individual rights. These rights are morally provided in the proviso. And the rights so grounded prove to be the familiar ones of our tradition—rights to person and to property. That they are also rationally grounded remains to be shown in section 5.

We must however recognize that these rights are not inherent in human nature. In defining persons for market competition and for co-operation, they assert the moral priority of the individual to society and its institutions. But they do not afford each individual an inherent moral status in relation to her fellows. In a pure state of nature, in which persons interact non-co-operatively and with no prospect of co-operation, they have no place. Rawls speaks of society as 'a cooperative venture for mutual advantage'.[21] It is only that prospect of mutual advantage which brings rights into play, as constraints on each person's behaviour. It is that prospect which enables rights to coexist with the assumption of mutual unconcern. The moral claims that each of us makes on others, and that are expressed in our rights, depend, neither on our affections for each other, nor on our rational or purposive capacities, as if these commanded inherent respect, but on our actual or potential partnership in activities that bring mutual benefit.

Each person acts rationally in seeking to maximize her utility, subject to two levels of constraint. First, each is constrained by the rights of her fellows, as determined by the proviso. If each respects

[20] Nozick, p. 9. [21] Rawls, p. 4.

the rights of others, then no one takes advantage of anyone else. The initial conditions for fair and rational market competition, and for the bargaining—explicit or implicit—that sets the terms for fair and rational co-operation, are satisfied. Second, each is constrained by the requirements of minimax relative concession, within co-operative institutions and practices. The error of some defenders of individual rights is to suppose that rights alone, the first level of constraint, are sufficient, and indeed, that further moral constraints on individual utility-maximization must be rejected as incompatible with rights. They would be correct if persons who adhered to the proviso found themselves always interacting with their fellows under conditions of perfect competition. For as we showed in Chapter IV, the market is a morally free zone. But the externalities that plague our interactions lead us to recognize that rights, although necessary, are morally insufficient. The compatibility of the two levels of constraint, respect for rights and adherence to co-operative practices, is a theme basic to the idea of morals by agreement.

5.1 Why should an individual seeking to maximize his utility not take advantage of his fellows? Why does his interest in market and co-operative arrangements, in itself an expression of his concern with utility-maximization, afford him reason to accept the proviso as limiting his rights in market and co-operative interaction? To bring our questions into sharper focus, let us consider once again the fisherfolk. You, the upstreamer, discharge your wastes into the running water of the river, thus causing pollution, and so costs for me, the downstreamer. This benefits you in interaction with me, and so brings the proviso into play; you lack the right to pollute.

There are two alternatives to consider—either your use of the river for waste disposal is the most efficient method overall, or it is not. If it is, then you should continue to pollute but compensate me for my resulting costs. If it is not, then you should adopt and pay for some other method of waste disposal that minimizes total costs. In each case optimality is achieved. But an optimal outcome may be realized in a quite different way. If using the river for waste disposal is efficient then you should simply continue to use it. If it is not, then I should pay you the difference in your costs necessary to induce you to adopt the most efficient method, since this payment must be less than the cost to me of your pollution. Again an optimal outcome is realized, although of course costs and benefits are distributed very differently between us.

When you and I agree to co-operate in a single society, you will no doubt contend that existing practices be continued, and I shall argue that those who create wastes should pay disposal costs. You will consider the pollution you bring about to be a simple consequence of natural interaction, and see no reason to refrain from it as a condition of co-operation—although you will refrain if I induce you to do so. I shall treat the effects of your pollution as a cost displaced on to me in natural interaction, and see no reason to accept it as a condition of co-operation—although I shall accept it if you compensate me. How may we resolve these opposed views? Appeal to the proviso would beg the question, since we are seeking its rationale. The proviso denies you, the upstreamer, the right to pollute, and awards me, the downstreamer, compensation should pollution occur. But do not the facts of the situation suggest a quite different assignment of rights? You do not coerce me by your method of waste disposal; you do not exact an unproductive transfer from me by predatory means. You simply dump your wastes in the river; if I object, then is not the onus on me to do something about it? The ban on coercion would seem to favour you, if anyone. Hence as a rational utility-maximizer you have no reason to change your method of waste disposal, and certainly no reason to make me a transfer payment, itself unproductive, as 'compensation'—unless I coerce you to pay. It may seem that reason and the proviso part company, and so, since the proviso ensures impartiality, do reason and morals.

But we deny this. Let us begin our rejoinder by recalling that co-operation has, as its sole and sufficient rationale, the maximization of expected utility. Thus in bargaining, the claim advanced and the concession offered by each person depend on his endeavour to maximize his utility, together with his recognition of the similar endeavour of every other person. The principle of minimax relative concession determines the outcome of co-operative interaction in such a way that shares in the co-operative surplus are related to contributions to its production in the same way for all. Of course, not every particular interaction, considered apart from a practice of co-operation, will benefit each party to it proportionately to his contribution. Everyone may expect to gain from certain arrangements for mutual assistance even though on any given occasion the recipient of assistance gains and the donor loses. It is the practice, and not the occasion, that must satisfy minimax relative concession.

Think now of the fisherfolk. I take a net loss if you dump your wastes in the river. Disposing of wastes by the method least costly to the disposer, ignoring all effects on others, is not a practice offering expected benefit to each member of society. The particular interaction cannot then be defended by relating it to a practice that satisfies minimax relative concession. Hence it violates the requirement, fundamental to rational co-operation, of mutual benefit proportionate to contribution.

If interaction is to be fully co-operative, it must proceed from an initial position in which costs are internalized, and so in which no person has the right to impose uncompensated costs on another. For if not, the resulting social arrangements must embody one-sided interactions benefiting some persons at cost to others. Even if each were to receive some portion of the co-operative surplus, yet each could not expect to benefit in the same relation to contribution as his fellows. Interactions based on displaced costs would be redistributive, and redistribution cannot be part of a rational system of co-operation.

But why should rational utility-maximizers interact in a fully co-operative manner? More generally, why should rational individuals enter fully into society, the locus of both market and co-operative interaction, rather than accepting particular market and co-operative practices within an enduring state of nature? Indeed, is not the state of nature the underlying reality of the human condition? These questions take us back to the position of the Foole. And here, as before, our answer to him turns on exhibiting the conditions for rational compliance with co-operative practices. In Chapter VI the Foole, appealing to the straightforward maximization appropriate to the state of nature, challenged the rationale for compliance with agreed joint strategies. There we showed that a rational individual will dispose himself to such compliance. Here the Foole challenges the rationale for limiting the effects of natural interaction in determining the individual rights from which social interaction proceeds. We must show that without limitations that exclude the taking of advantage, a rational individual would not dispose himself to co-operative compliance.

In VI.2.4 we distinguished broad and narrow compliance. A person disposed to broad compliance compares the benefit she would expect from co-operation on whatever terms are offered with what she would expect from non-co-operation, and complies if the

former is greater. Were persons so disposed, then no one would have reason to accept the proviso, or any other constraint, on natural interaction. The non-co-operative outcome would serve as base-point, and any improvement on it would elicit voluntary co-operation. But broad compliance is not a rational disposition for utility-maximizers. Not only does a broadly compliant person invite others to take advantage of her in setting terms of co-operation, but if some persons are broadly compliant, then others, interacting with them, will find it advantageous not to be broadly, or even so much as narrowly, compliant. If you will comply for any benefit whatsoever, then in interacting with you I should dispose myself to comply with a joint strategy only if it offers me, not a fair share, but the lion's share of the co-operative surplus. So it is not and cannot be rational for everyone to be disposed to broad compliance. But since no one chooses to constrain his behaviour for its own sake, no person finds it rational to be more compliant than his fellows. Equal rationality demands equal compliance. Since broad compliance is not rational for everyone, it is not rational for anyone.

A person disposed to narrow compliance compares the benefit he would expect from co-operation with what he would expect from a fair and optimal outcome, and complies with a joint strategy only if the former approaches the latter. An outcome is fair in satisfying the standards of impartiality, which as we have shown are set by the proviso and the principle of minimax relative concession. Thus a person disposed to narrow compliance expects others to adhere, and to consider it rational to adhere, to the proviso as a condition of co-operation. But then, given equal rationality, he must consider it rational to adhere himself to the proviso as a condition of co-operation. The disposition to narrow compliance thus includes the disposition to accept the proviso as constraining natural interaction, in so far as one has the expectation of entering into society, into market and co-operative practices.

The rationality of disposing oneself to narrow compliance, and so to acceptance of the proviso, follows from the advantageousness of society and the equal rationality of its members. If all persons are less than narrowly compliant, refusing to act voluntarily on joint strategies leading to fair and optimal outcomes, then co-operation is not possible. If some persons are less than narrowly compliant, then co-operation is possible only if others are more compliant. But this violates equal rationality. If some persons are more than narrowly

compliant, then others would find it advantageous, and so rational, to be less compliant. But this again violates equal rationality. It is rational for each person to be sufficiently compliant that society is possible if others are equally compliant; it is not rational for anyone to be so compliant that society is possible if others are less compliant; therefore it is rational for each person to be narrowly compliant.

Each person endeavours to maximize his own utility. In so doing, each seeks to take what benefits he can from the actions of others, and to displace what costs he can from his own actions to theirs. But also, each seeks to avoid interactions with others that afford them benefits for which he pays the costs. In purely natural interaction, the desire to avoid costly interaction has no independent force to constrain the desire to benefit oneself without regard to others. But in social interaction, requiring voluntary constraint, the desire to avoid costly interaction can be balanced with the desire to take benefits and displace costs. The proviso, forbidding the taking of advantage, represents the weakest constraint rationally acceptable to persons who would avoid costly interaction with others, and the strongest constraint rationally acceptable to persons who would be free to benefit themselves. Thus the proviso reflects the equal rationality of persons who must constrain their natural interaction in order to enter into mutually beneficial social relationships.

As we have seen, interaction constrained by the proviso generates a set of rights for each person, which he brings to the bargaining table of society as his initial endowment. He brings a right to his person, a right in the fruits of his labour, and a right to those goods, whose exclusive individual possession is mutually beneficial, that he has acquired either initially or through exchange. Where goods are not suitable for or capable of such possession, he brings a right of use on terms that internalize all costs. Without these rights, persons would not be rationally disposed, either to accept the prohibition on force and fraud needed for market competition, or to comply voluntarily with the joint strategies and practices needed for co-operation.

5.2 But is our reconciliation of reason and morals not too good to be true? Recall the tale of masters and slaves with which this chapter began. We introduce here Fig. 9, intepreted as illustrating the story, and Fig. 10, representing a significant variant of the tale. In both figures, I_n corresponds to the initial situation of masters

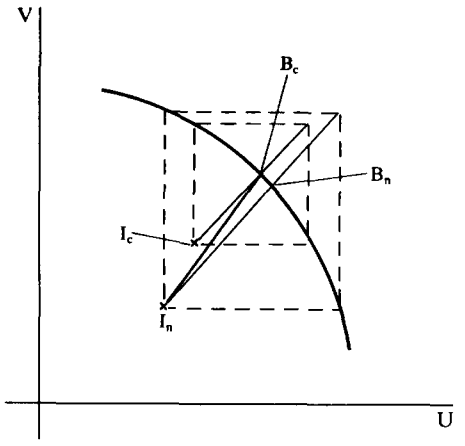

Figure 9

and slaves, B_n to the outcome of agreement as imagined by the young master, in which the ex-slaves voluntarily continue to serve, B_c to the real outcome in which the ex-slaves refuse to comply with the unproductive transfer involved in continued service, and I_c to the hypothetical situation, non-co-operative but non-coercive, that is determined by the proviso as the initial position for bargaining. The difference between the two variants of the tale is evident. If events occur as illustrated in Fig. 10, the final outcome for the ex-masters, represented by B_c, affords them a lesser expectation of utility than their initial situation, I_n. They lose by co-operation, if the initial bargaining position is constrained by the proviso.

Surely the masters, if rational and able to foresee the outcome, would never agree to co-operate on terms that would prove disadvantageous to them. We cannot suppose that Fig. 10 illustrates a real possibility. Rather, we must surely suppose either that no co-operation takes place, so that masters and slaves remain at I_n, or that co-operation takes place on terms affording mutual benefit to both, leading to a point such as B_a in Fig. 10, where the ex-slaves continue to serve the ex-masters to an extent sufficient to make the masters better off than at I_n.

Neither alternative seems palatable. The first enables us to hold fast to our insistence that it is rational to dispose oneself to comply

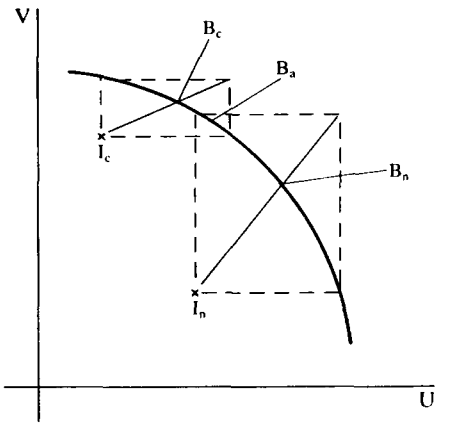

Figure 10

with co-operative arrangements only if they are (nearly) fair and optimal. Then the slaves are unwilling to accept any outcome that affords some portion of the co-operative surplus to the masters. We must deny that there can be any rational improvement on the sub-optimal non-co-operative outcome. The second alternative requires us to weaken the proviso and strengthen the disposition to comply. The masters' rights in the initial bargaining position allow them to take some advantage of the slaves. The slaves are willing to comply with a joint strategy leading to an optimal outcome that is as fair as is consistent with affording both parties mutual benefit. The link between rationality and impartiality is broken. (A third alternative, that the masters should accept the constraints of the proviso and co-operate even though they lose in relation to non-co-operation, we dismiss as clearly incompatible with a utility-maximizing conception of rationality, however constrained.)

For utility-maximizers, the link between co-operation and mutual benefit must take precedence over the link between co-operation and impartiality or fairness. But this does not require us to sever the second link. The proviso constrains the initial bargaining position to the extent, but only to the extent, that such constraint is compatible with the co-operative outcome affording each person the expectation of a utility greater than that afforded by the non-co-operative

outcome. It is rational to comply with a co-operative joint strategy if and only if its expected outcome is (nearly) optimal and as fair as is compatible with mutual benefit. We abandon neither the proviso nor narrow compliance, but we subordinate them to the requirement of mutual benefit.

We should distinguish between compliance, as the disposition to accept fair and optimal co-operative arrangements, and *acquiescence*, the disposition to accept co-operative arrangements that are less than fair, in order to ensure mutual benefit. A person who acquiesces in a joint strategy does not consider her own greatest utility on the particular occasions when she follows the strategy. She constrains her utility-maximizing endeavours. But she does not ignore the advantage that she concedes to others in acquiescence, and she remains ready to withhold it, and to demand a joint strategy more favourable to herself, if she supposes that such a strategy would not make co-operation unprofitable to others. Here we suppose that each person does appeal to the natural distribution or non-co-operative outcome which underlies existing social realities, as Buchanan argues.[22] As this underlying equilibrium changes, the conditions of acquiescence also change.

Co-operation on terms less than fair is therefore less stable, in failing to gain the whole-hearted acceptance of all participants. Fair co-operation invites a full compliance which each does not stand ready to withdraw because of shifts in the natural distribution. For although no one is prepared to concede advantage to others in order to bring about this stability, each is willing to accept it, as enhancing the benefits of co-operative institutions and practices, when it is not costly to herself to do so.

We suppose that the link between reason and morals is loosened because of the role that three quite different considerations play in affecting interaction. First, there are ideological factors—beliefs that affect the terms of co-operation in ways that are unfair, sometimes to those who share the beliefs but typically to those who do not. If it is generally held that a woman's place is in the home, then a woman who does not find the home-maker's role satisfying may find it impossible to interact with others in a way that affords her a fair share of the co-operative surplus. Second, and often linked with ideology, are historical factors—institutions and practices that effectively determine the rights persons can exercise in interaction, whether these satisfy the proviso or not. If it is customary to pay women lower wages than men, no particular woman may reasonably expect to be recognized as entitled to equal pay for equal work.

[22] See Buchanan, pp. 74–82, esp. p. 79.

These factors may, ultimately, depend on irrational beliefs—although the irrationality may be strongly ingrained in social practices and institutions that individuals are not effectively able to ignore. But this irrationality makes them less disturbing to our argument linking reason and morals than the third, technological factor. Technology is power; those with a more advanced technology are frequently in a position to dictate the terms of interaction to their fellows. Without their guns, a small number of Spaniards would never have been able to overcome the Indian civilizations of the Americas. To be sure, the Spaniards preferred to dominate rather than to co-operate, but it is clear that their technological superiority ensured that any co-operation must have proceeded from an initial position in violation of the proviso.

A superior technology enables its possessors rationally to maintain, and requires others rationally to acquiesce in, arrangements that rest on differential rights in clear violation of the proviso. And we may not suppose, as with ideology and its institutionalized effects, that the basis of this technology is irrational. On the contrary, it constitutes the most fully rational application of belief to practice. We may say that those possessing a superior technology are more rational than their fellows, in being better able to relate and devise means to their ends. And this reveals why technological factors weaken the link between rationality and impartiality. Our argument in support of the proviso, and in support of narrow compliance, rests on an assumption of equal rationality among persons which differences in technology deny.

Nevertheless, we suppose that the unequal rationality brought about by technological differences between societies is accidental. It does not reflect underlying differences in the capacities of the persons constituting those societies. Freed from false views of the world and the practices and institutions to which such false views give rise, human beings tend toward technological equality. They tend, then, toward a state of affairs in which the proviso, the principle of minimax relative concession, and the disposition to narrow compliance, come fully into play in linking the requirements of reason to the demands of impartiality or morals.

In reconciling reason and morals, we do not claim that it is never rational for one person to take advantage of another, never rational to ignore the proviso, never rational to comply with unfair practices. Such a claim would be false. We do claim that justice, the disposition not to take advantage of one's fellows, is the virtue appropriate to co-operation, voluntarily accepted by equally rational persons. Morals arise in and from the rational agreement of equals.

VIII

THE ARCHIMEDEAN POINT

1.1 'Embedded in the principles of justice there is an ideal of the person that provides an Archimedean point for judging the basic structure of society.'[1] Moral theory offers an Archimedean point analysis of human interaction. But what is an Archimedean point? The reader will recall that Archimedes supposed that given a sufficiently long lever and a place to stand, he could move the earth. We may then think of an Archimedean point as one from which a single individual may exert the force required to move or to affect some object. In moral theory, the Archimedean point is that position one must occupy, if one's own decisions are to possess the moral force needed to govern the moral realm. From the Archimedean point one has the moral capacity to shape society. We characterize the point in terms of its occupant, for it is the characteristics intrinsic to a person, and not her external contrivances or circumstances, which afford her moral force.

What ideal of the person does the Archimedean point express? How does that ideal lead to a judgement of the structure of society, or of the basic principles that underlie social interaction? And what is the import of this judgement for reason and for morals? Briefly, we suppose that the ideal presents a rational actor freed, not from individuality but from the content of any particular individuality, an actor aware that she is an individual with capacities and preferences both particular in themselves and distinctive in relation to those of her fellows, but unaware of which capacities, which preferences. Such a person must exhibit concern about her interactions with others, and this concern leads her to a choice among possible social structures. But her concern is necessarily impartial, because it is based on the formal features of individual rational agency without the biasing content of a particular and determinant set of individual characteristics.

We may think of the choice of this ideal person as proceeding at two levels. Abstractly it is a choice among principles of interaction; concretely it is a choice among social structures embodying these

[1] J. Rawls, *A Theory of Justice* (Cambridge, Mass., 1971), p. 584.