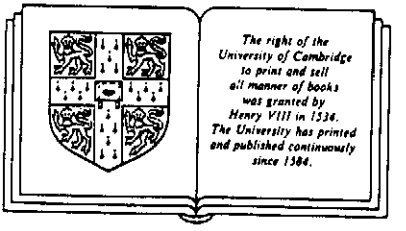
Contractarianism and Rational Choice

Essays on David Gauthier's *Morals by Agreement*

Peter Vallentyne, editor



Cambridge University Press

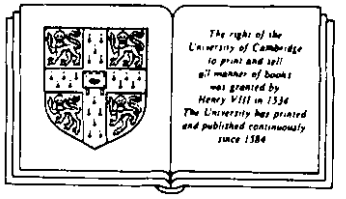
New York

Cambridge Port Chester Melbourne Sydney

Contractarianism and Rational Choice

Essays on David Gauthier's *Morals by Agreement*

Peter Vallentyne, editor



Cambridge University Press

New York

Cambridge Port Chester Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1991

Printed in the United States of America

Library Cataloging in Publication Data
Contractarianism and rational choice : essays on David Gauthier's Morals by agreement / Peter Vallentyne, editor.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-39134-2.—ISBN 0-521-39815-0 (pbk.)
1. Gauthier, David P. Morals by agreement. 2. Ethics. 3. Contracts. 4. Cooperation. I. Vallentyne, Peter
BJ1012.038 1991
171—dc20 90-387744
CIP

British Library Cataloguing in Publication Data
Contractarianism and rational choice : essays on David Gauthier's Morals by agreement.
1. Moral philosophy. Theories of Gauthier, David P. (David Peter)
I. Vallentyne, Peter
170.92

ISBN 0-521-39134-2
ISBN 0-521-39815-0 pbk

Contents

Preface

This anthology started out as a collection of papers presented at a conference on contemporary contractarian thought that I organized in April 1987 at the University of Western Ontario. Not surprisingly, all but a few of the essays dealt primarily with the work of David Gauthier. In order to increase the focus and coherence of the volume, I decided to drop the pieces that were not specifically on Gauthier's work and to add some pieces on aspects of Gauthier's project not addressed by the other papers.

A great number of thanks are due. To start, the conference from which this volume comes, and some of the prepublication costs, were funded by a grant from the Social Sciences and Humanities Research Council of Canada (#443–87–0012). The conference was also funded by grants from The Faculty of Arts (via the Philosophy Department), The Faculty of Social Sciences, and The Center for Critical Theory at The University of Western Ontario. Terence Moore, Humanities Editor at Cambridge University Press, has been a true delight to work with. He has been extremely supportive and reasonable. David Gauthier kindly agreed to find the time in a tight schedule to write a concluding comment for the volume. David Copp, Jean Hampton, and Chris Morris each provided important advice about getting the volume published. Most of all, however, I owe a deep debt of gratitude to Jules Coleman and Geoff Sayre-McCord. Right from the beginning, Geoff was my main consultant, and without his support and advice, the volume would not have made it to press. Jules Coleman was not involved in the beginning stages, but he was absolutely crucial at the later stages. He very kindly took the time on several occasions to provide specific suggestions on how the volume could be improved in content and organization. Finally, all the contributors have been extremely patient with the slow pace at which the volume has come togetber.

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1991

Printed in the United States of America

Library Cataloging in Publication Data
Contractarianism and rational choice : essays on David
Gauthier's Morals by agreement / Peter Vallentyne, editor.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-39134-2.—ISBN 0-521-39815-0 (pbk.)
1. Gauthier, David P. Morals by agreement. 2. Ethics.
3. Contracts. 4. Cooperation. I. Vallentyne, Peter
BJ1012.038 1991
171—dc20 90-387744
CIP

British Library Cataloguing in Publication Data
Contractarianism and rational choice : essays on David
Gauthier's Morals by agreement.
1. Moral philosophy. Theories of Gauthier, David P. (David Peter)
I. Vallentyne, Peter
170.92

ISBN 0-521-39134-2
ISBN 0-521-39815-0 pbk

Contents

Preface

This anthology started out as a collection of papers presented at a conference on contemporary contractarian thought that I organized in April 1987 at the University of Western Ontario. Not surprisingly, all but a few of the essays dealt primarily with the work of David Gauthier. In order to increase the focus and coherence of the volume, I decided to drop the pieces that were not specifically on Gauthier's work and to add some pieces on aspects of Gauthier's project not addressed by the other papers.

A great number of thanks are due. To start, the conference from which this volume comes, and some of the prepublication costs, were funded by a grant from the Social Sciences and Humanities Research Council of Canada (#443–87–0012). The conference was also funded by grants from The Faculty of Arts (via the Philosophy Department), The Faculty of Social Sciences, and The Center for Critical Theory at The University of Western Ontario. Terence Moore, Humanities Editor at Cambridge University Press, has been a true delight to work with. He has been extremely supportive and reasonable. David Gauthier kindly agreed to find the time in a tight schedule to write a concluding comment for the volume. David Copp, Jean Hampton, and Chris Morris each provided important advice about getting the volume published. Most of all, however, I owe a deep debt of gratitude to Jules Coleman and Geoff Sayre-McCord. Right from the beginning, Geoff was my main consultant, and without his support and advice, the volume would not have made it to press. Jules Coleman was not involved in the beginning stages, but he was absolutely crucial at the later stages. He very kindly took the time on several occasions to provide specific suggestions on how the volume could be improved in content and organization. Finally, all the contributors have been extremely patient with the slow pace at which the volume has come together.

of trolley cars while still in his pram, is in what is now called light rail transit. When much younger, he was an unsuccessful candidate for election to the Canadian House of Commons, an occasional newspaper columnist, and a writer on public affairs.

1. Gauthier's three projects

Peter Vallentyne

Introduction

Over the last twenty years, and most notably in his recent book *Morals by Agreement*, David Gauthier has been engaged in three distinct, but closely related, projects: (1) defending a contractarian theory of morality, (2) defending a theory of rational choice, and (3) defending the claim that rationality requires that we comply with the dictates of morality.[1] In this introduction, I give a brief overview of Gauthier's work (as it appears in his book) on each of these three projects. We start with his views on morality.

Gauthier's moral methodology

In a world in which the conditions of perfect competition hold – most importantly: there are no externalities, that is, no one is affected by the actions of others except by consent – the unconstrained pursuit of self-interest leads to Pareto optimal results (i.e., results such that no one can be made better off except by making someone else worse off). In such a world, there are, Gauthier holds, no moral constraints on action. In

[1] Some of the main articles leading up to Gauthier's book *Morals by Agreement* (Oxford: University Press, 1986) are "Morality and Advantage," *Philosophical Review* 76 (1967): 460–75; "Rational Cooperation," *Noûs* 8 (1974): 53–65; "Reason and Maximization," *Canadian Journal of Philosophy* 4 (1975): 411–33; "The Social Contract as Ideology," *Philosophy and Public Affairs* 6 (1977): 130–64; and "The Social Contract: Individual Decision or Collective Bargain?" in *Foundations and Applications of Decision Theory*, Vol. 2, edited by Cliff Hooker, Jim Leach, and Edward McClennen (Dordrecht, Holland: D. Reidel, 1978), pp. 47–67. Since the publication of the book, a number of journals have devoted volumes to a critical assessment of his work. See, for example, the articles in *Ethics* 97(4) (1987), *Canadian Journal of Philosophy* 18(2) (1988), and *Social Philosophy and Policy* 5(2) (1988).

our world, however, the conditions of perfect competition do not hold. In particular, the world is full of externalities. Consequently, as the familiar Prisoner's Dilemma illustrates, the unconstrained pursuit of self-interest often leads to outcomes that are to the disadvantage of all. The role of morality, in Gauthier's view, is to constrain the pursuit of self-interest so as to ensure Pareto optimal results.

Gauthier explicitly equates moral constraint with rational and impartial constraint on the pursuit of self-interest. Although most people would accept that being a rational and impartial constraint on conduct is a necessary condition for being a moral constraint, many would reject it as a sufficient condition. One might claim, for example, that an element of sympathy for others is also necessary.

Gauthier is not, however, very interested in arguing about the proper conception of morality. His main interest is to give an account of rational and impartial constraints on conduct. If this does not capture the traditional conception of morality, so much the worse for the traditional conception. Rationality – not morality – is the important notion for him.

Gauthier's lack of concern for the traditional conception of morality is apparent in his rejection of any appeal to moral intuitions. He writes:

> If the reader is tempted to object to some part of this view, on the ground that his moral intuitions are violated, then he should ask what weight such an objection can have, if morality is to fit within the domain of rational choice. (*Morals by Agreement*, p. 269)

Gauthier rejects not only brute appeal to moral intuition, but also appeal to our considered moral judgments in reflective equilibrium.[2] Consequently, his project is best understood as potentially involving a radically reformist conception of morality. It is not merely that his theory might fail to capture some traditional moral concerns, but rather that its connection with these traditional moral concerns is purely contingent.[3]

Gauthier's concern with rationally acceptable norms of interaction leads him to advocate a contractarian theory quite unlike that of many contractarians, and, in particular, quite unlike that of Rawls. For unlike Gauthier, Rawls' project is not to reduce morality to rationality (and disregard the rest), but rather to use the theory of rational choice to derive moral principles from a *morally loaded* choice situation. Rawls' original position (and his veil of ignorance in particular) is set up to screen *morally irrelevant* features of the status quo (e.g., one's position and capacities). Gauthier, on the other hand, wants to use the theory

[2] See Norman Daniels, "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *Journal of Philosophy* 76 (1979): 256–82, for an excellent discussion and elaboration of the method of reflective equilibrium.

[3] In his essay in this volume (chap. 1), Gauthier further elaborates on his moral methodology.

of rational choice to derive moral principles from a *morally neutral* choic situation. For both theories, the relevant agreement is hypothetical, bu the relevant circumstances of agreement are grounded, we shall see much more closely in reality on Gauthier's view than on Rawls'.

Gauthier's contractarian theory

Contractarian moral theories hold that an action, practice, law, or soci structure is morally permissible just in case it, or principles to which conforms, would be (or has been) agreed to by the members of societ under certain specified conditions.[4] Like most contemporary contrac tarian theorists, Gauthier takes the relevant agreement to be hypothetica (i.e., what *would* be agreed to under appropriate circumstances) – nc actual agreement.[5]

Gauthier advocates both a contractarian ethical theory and a contrac tarian political theory.[6] His contractarian ethical theory is *indirect* in tha it judges actions permissible if and only if *they conform to principles* tha would be chosen under specified conditions. His contractarian politica theory, on the other hand, is *direct* in that it judges social structure permissible if and only if *they* (as opposed to principles to which the conform) would be chosen under the specified conditions.[7] Thus, on

[4] Contractarian moral theory has been around at least since it was very briefly suggeste by Glaucon in Book II of Plato's *Republic*. The main historical works are Thomas Hobbes *Leviathan* (1651), edited by C. B. Macpherson (New York: Penguin, 1968); John Locke *The Second Treatise of Government* (1690), edited by Thomas Peardon (Indianapolis: Bobbs Merrill, 1952); Jean-Jacques Rousseau, *Contrat Social* (1762), in *Political Writings*, edite by C. E. Vaughan (Cambridge: Cambridge University Press, 1915); Immanuel Kant *Groundwork of the Metaphysics of Morals* (1785), edited by H. J. Patton (New York: Harper 1958); and Immanuel Kant, *The Metaphysical Elements of Justice* (1797), translated an edited by John Ladd (Indianapolis: Bobbs-Merrill, 1965). The main contemporary wor is, of course, John Rawls, *A Theory of Justice* (Cambridge, MA: Belknap Press of Harvar University Press, 1971). In addition to Gauthier's book, some important recent work are James Buchanan, *The Limits of Liberty* (Chicago: The University of Chicago Press 1975); T. M. Scanlon, "Contractualism and Utilitarianism," in *Utilitarianism and Beyond* edited by Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press 1982); Gilbert Harman, "Justice and Moral Bargaining," *Social Philosophy and Policy* (1983): 114–31; Jean Hampton, *Hobbes and the Social Contract Tradition* (New York: Cam bridge University Press, 1986); and Gregory Kavka, *Hobbesian Moral and Political Theor* (Princeton: Princeton University Press, 1986).

[5] One of the few authors to advocate actual agreement contractarianism is Gilbert Harman "Justice and Moral Bargaining," *Social Philosophy and Policy* 1 (1983): 114–31.

[6] That Gauthier advocates a political as well as an ethical contractarian theory is evidence by his discussion in chap. 8 of choosing social structures. Note that in "Contractualism and Utilitarianism," T. M. Scanlon advocates a contractarian ethical theory, whereas John Rawls, in *A Theory of Justice*, and James Buchanan, in *The Limits of Liberty*, advocate (primarily) contractarian political theories.

[7] The distinction between direct and indirect contractarian theories parallels that between direct and indirect (e.g., act vs. rule) utilitarian theories. The difference lies in whether the object of potential agreement is the object of which the permissibility is being assessed (e.g., an action) or whether it is a set of principles that will assess the object being assessed.

way in which Gauthier's political theory differs from that of Rawls is that he has the parties agreeing on particular social structures, whereas Rawls has them agreeing on principles for assessing social structures.

Gauthier's specification of the conditions under which the agreement takes place is also quite different from that of Rawls. Let us consider his specification of (1) the parties to the agreement, (2) the beliefs of the parties, and (3) the desires of the parties.

The parties to the agreement

Gauthier takes the parties to the agreement for a given society at a given time to include only those currently living members of society whose cooperation would benefit at least some other currently living members of society. Gauthier explicitly takes the hard line that animals, children, the severely disabled, and members of future generations are *not* parties to the agreement, since they offer no benefit to current, nondisabled, adult human beings beyond what the latter can obtain unilaterally.[8] Gauthier's project is to ground morality in rational agreement, and rational agreement, he maintains, requires mutual advantage. Consequently, the protection of those too weak to offer benefits to others is, in Gauthier's view, included in the scope of morality only to the extent that those party to the agreement care about them.

The beliefs of the parties

Unlike Rawls' theory, which imposes a thick veil of ignorance on the parties (allowing them only knowledge of the laws of nature and of some very general features of the state of the world), Gauthier's theory imposes no veil of ignorance at all. Gauthier's social agreement is based on rational negotiation among fully informed, determinate individuals.

Allowing the parties to be fully informed about their capacities and situations allows the parties with more advantageous capacities and positions to bargain for a greater share of the benefits. In Rawls' original position, on the other hand, advantaged parties have no information about their capacities and positions, and, therefore, have no basis for bargaining for a greater share of the benefits. Here again Gauthier's project of reducing morality to rationality leads him to take a tough-minded position regarding the weak. Rationality requires that all available information be used, and since that leads to weaker bargaining

[8] In "Contractualism and Utilitarianism," Scanlon, on the other hand, allows children, animals, and future people to be represented in the agreement by a trustee.

positions for the weak, it is unreasonable for the weak to expect the same benefits as the strong.[9]

The desires of the parties

Like Rawls, Gauthier assumes that the parties are mutually unconcerned (do not care how others fare).[10] Since people do, at least to some extent, care – both negatively and positively – about how other people's conceptions of the good are promoted, this assumption is counterfactual. It seems, therefore, to detract from the rational grounding of the resulting agreement. An agreement that would be rational if people had certain sorts of desires need not be (and in general is not) rational, if those people do not have those desires. This is especially problematic for Gauthier, since (unlike Rawls, who could argue that other-regarding desires are morally irrelevant to the choice of principles), Gauthier wants to ground his theory solely in considerations of rational choice. He does not want to rely on any assumptions about what is morally relevant. The rationality of an agreement based on the assumption that people do not have other-regarding desires does not show that such an agreement is rational for people with such desires. So this may be a problematic feature of Gauthier's theory.[11]

These, then, are Gauthier's views on morality. Let us now consider his views on rational choice.

Gauthier's rational-choice framework

Gauthier defends an instrumental conception of practical rationality, according to which a choice is rational if and only if relative to the agent's beliefs it is the most effective means for achieving the agent's goals. Except for certain minimal formal conditions of coherence,[12] the instru-

[9] We should note here an alternative approach similar in spirit to Rawls'. Both Thomas Scanlon, "Contractualism and Utilitarianism," and Jürgen Habermas, *Legitimation Crisis*, translated by Thomas McCarthy (Boston: Beacon Press, 1975), have advocated contractarian theories that (like Gauthier's theory) allow the parties full knowledge concerning their situations, capacities, beliefs, and desires, but prevent the parties from simply pursuing their own self-interest, by specifying that the parties are motivated solely by the desire to reach a reasonable agreement. The agreement of these contractarian theories is thus to be understood as that of a consensus rather than a bargain or compromise. Rawls' theory is also consensualist, since the veil is so thick there is no room for conflict and compromise. What Rawls achieves through a veil of ignorance, these authors achieve through a special motivation assumption.

[10] Note that the assumption of mutual unconcern applies only to the parties in the choice situation (in which they come to an agreement). The people whom the chosen principles are to regulate are not assumed to be mutually unconcerned.

[11] I develop this point in chap. 5, "Contractarianism and the Assumption of Mutual Unconcern," in this volume.

[12] Such as being considered, complete, transitive, monotonic in prizes, and continuous.

mental conception of rational choice rejects any attempt to assess the rationality of the goals themselves. Value (utility), Gauthier argues, is subjective (dependent on the affective attitudes of individuals) and relative (not necessarily the same for all individuals). There are no external norms for assessing someone's preferences, Gauthier claims, except the formal coherence properties.

In *parametric choice* – that is, in choice situations in which agents take their environment as fixed – a choice is rational if and only if it maximizes expected utility (i.e., is the most effective means for achieving the agent's goals). In *strategic choice* – that is, in choice situations in which the agent recognizes that the outcome of choice depends in part on the choices of other rational agents – the rationality of a choice depends (in part) on what it would be rational for the agents to agree to. What, then, determines whether an agreement is rational?

The problem of rational agreement is to select a single option (or perhaps a set of options) from a set of feasible options in a way that is rationally acceptable to all the parties to the agreement. On the received view, which Gauthier accepts, rational agreement can be reconstructed as a two-step process. In the first step, an initial bargaining position is determined. This position determines the utility payoff that each person brings to the bargaining table and that is not subject to negotiation. It is only the utility payoff over and above the initial bargaining position payoff that is negotiable. In the second step, an option (or set of options) is chosen on the basis of the initial bargaining position. As we shall now see, Gauthier has a new and interesting account of both steps.

The initial bargaining position

A common specification of the initial bargaining position is as the *noncooperative outcome.*[13] This is the hypothetical outcome of the uncoordinated pursuit of self-interest. If agreement is based on the noncooperative outcome, then, although the benefits of cooperation over noncooperation will be distributed among the parties, the distribution of the costs and benefits of noncooperation will be untouched. Agents will end up with the net benefits that they would obtain from noncooperation and a portion of the benefits of engaging in cooperative behavior.

Gauthier agrees that a rational agreement should make everyone at least as well off as under the noncooperative outcome, but argues nonetheless that the noncooperative outcome should not be treated as the initial bargaining position. A rational agreement, Gauthier claims, must

[13] See, for example, James M. Buchanan, *The Limits of Liberty: Between Anarchy and Leviathan* (Chicago: The University of Chicago Press, 1975) for a defense of the noncooperative outcome as the initial bargaining position.

be one with which it is rational for the agents to comply, and agreements based on the noncooperative outcome do not satisfy that condition. In particular, it is irrational, he argues, for those who would be net victims of noncooperative interaction (i.e., those who would be worse off in the presence of the noncooperative activities of others than they would be if left completely alone) to comply with agreements based on the noncooperative outcome. Such agreements would perpetuate the benefits and costs of coercive activity even though such activity would no longer take place, and therefore be unstable.

Gauthier claims that, in order for there to be a rational basis for all to comply with an agreement, the initial bargaining position must be the hypothetical result of noncooperative interaction *constrained by the Lockean Proviso,* that is, of noncooperative interaction subject to the constraint that no one makes him/herself better off by making someone else worse off.[14] Like the noncooperative outcome, this represents an outcome where all social cooperation ceases. Unlike the noncooperative outcome, however, it is based on the assumption that no one engages in coercive or predatory activity. People neither help nor harm others.

Many will be unconvinced by Gauthier's defense of the relevance of the proviso for the theory of rational choice. The initial bargaining position must, they will argue, reflect how people would fare in the absence of cooperation. And in the absence of cooperation, it is sometimes rational for individuals to better their own position by worsening that of others. The appropriate baseline (the objection continues) is simply the noncooperative outcome – not that outcome constrained by the proviso.

Gauthier's specification of the initial bargaining position for rational agreements is thus both novel and controversial. As we shall now see, his views on the rational allocation of the benefits of cooperation (i.e., the benefits beyond those of the initial bargaining position) are also novel and controversial.

The bargaining solution

The bargaining problem is this: Given a set of feasible options, one of them being the initial bargaining position, which option(s) is it rational

[14] On p. 212, Gauthier makes clear that the proviso is weaker than it might seem. It only rules out making someone worse off when worsening the situation of others is *the means* – as opposed to a mere side effect – of bettering one's own situation. Thus, taking without your consent the fish you have caught is prohibited, since I improve my lot *by* worsening your lot; but polluting the river (and thereby killing many of the fish you might otherwise catch) does *not* violate the proviso, since the fact that you are made worse off is purely incidental to the benefit I get from polluting the river (I would still get the benefit, if you did not exist). This raises two questions: Can a clear distinction between means and side effects really be made? If it can, is Gauthier's interpretation of the proviso really rationally more acceptable than a version that prohibits bettering one's situation by worsening – as means or as a side effect – another's situation?

to choose? A bargaining solution is a specification of a procedure for answering this question. The most well-known solution is the Nash solution (also known as the Zeuthen–Nash–Harsanyi solution), according to which rational agreement fixes on an option that maximizes the (mathematical) product of each person's excess utility over the initial bargaining position. Thus, for each feasible option O one calculates the value $[U_1(O) - U_1(O^*)] \times [U_2(O) - U_2(O^*)] \times \cdots \times [U_n(O) - U_n(O^*)]$, where O^* is the initial bargaining position point, and U_i is person i's utility function (where i is $1, 2, \ldots, n$). According to the Nash solution, a rational agreement maximizes this product. For example, if there are two agents and four options, ⟨0,0⟩, ⟨100,30⟩, ⟨50,50⟩, and ⟨0,100⟩,[15] with ⟨0,0⟩ being the initial bargaining position, then it is rational to agree upon ⟨100,30⟩ (since $(100 - 0) \times (30 - 0)$ [= 3,000] is greater, for example, than $(50 - 0) \times (50 - 0)$ [= 2,500]).

Gauthier defends a different solution. He claims that rational agents would choose a feasible option that *minimizes the maximum relative concession* that anyone makes. The *relative concession* that a person makes for a given option is the ratio of (a) the excess of (i) the utility for that person of his/her most favorable admissible option over (ii) the utility for that person of the given option to (b) the excess of (i) the utility for that person of his/her most favorable admissible option over (ii) the utility for that person of the initial bargaining position option. An admissible option is one that is both feasible and accords everyone at least as much utility as the initial bargaining position point.[16] In symbols, the relative concession for person i of option O is $[U_i(O_i) - U_i(O)]/[U_i(O_i) - U(O^*)]$, where O^* is the initial bargaining position option, and O_i is i's most favored admissible option (i.e., the admissible option that gives i as at least as much utility as any other admissible option).[17] In the example choice situation of the preceding paragraph, it is rational according to Gauthier to agree upon ⟨50,50⟩, since its maximum relative concession is 0.5 (50/100), and, for example, the maximum relative concession of ⟨100,30⟩ is 0.7 (70/100).

According to Gauthier's bargaining solution, then, rational agents would agree to an option for which the highest relative concession is as low as possible. Very roughly, the intuitive idea behind this solution is

[15] ⟨n,m⟩ represents an option that yields n units of utility to player one and m units to player two.

[16] The minimax relative concession principle is also known as the maximin relative benefit principle, where relative benefit is the proportion received of one's maximum admissible benefit over the initial bargaining position. Minimizing the maximum relative concession is equivalent to maximizing the minimum relative benefit.

[17] Neither Nash's solution nor Gauthier's requires that utility be interpersonally comparable. Because both involve utility differences (e.g., $U_i(O_i) - U_i(O^*)$), they do not require that the zero points of different people's utility scales be comparable. Because (unlike utilitarianism) neither adds one person's utility with that of another, they do not require that the units of different people's scales be comparable.

that, since one's ground of complaint can be measured by one's relative concession, minimizing maximum relative concession minimizes the grounds for complaint.

Both Nash's and Gauthier's solutions have been axiomatized, so the differences between the two can be traced back to differences in which axioms are accepted and are rejected.[18] The difference lies in Nash's acceptance, and Gauthier's rejection, of Condition Alpha:[19] if an option is a rational choice for a given set of feasible options and given initial bargaining position point, then it is also a rational choice for any *subset* of these options with the same initial bargaining position point. The idea is that an option should not cease to be a rational choice simply because some of its competitors are eliminated.

Gauthier rejects Condition Alpha because he holds that what is rational for one to accept depends on how favorable is one's most favorable admissible option. Thus, he holds that an option that is a rational choice for a given initial bargaining position point and a given set of feasible options may not be a rational choice given the same initial bargaining position point and a *subset* of those feasible options. Indeed, Gauthier holds that, in general, such an option will *not* be a rational choice for the subset, if the subset was obtained by eliminating someone's most favorable admissible option.[20]

Given the plausibility of Gauthier's claim about the relevance of options that are someone's most favorable admissible option, his bargaining solution represents an important challenge to the status of Nash's solution as the received solution.

The rationality of complying with rational agreements

The combination of Gauthier's specification of the initial bargaining position (the hypothetical outcome of noncooperative interaction constrained by the Lockean Proviso) and his bargaining solution (minimize the maximum relative concession) specifies what it is rational for agents

[18] See, for example, the discussion of the Kalai and Smordinsky solution (solution G) in Alvin Roth, *Axiomatic Models of Bargaining* (New York: Springer-Verlag, 1979). Gauthier's solution is only slightly different from that solution. The two were arrived at independently.

[19] Condition Alpha is also known as the Independence of Irrelevant Alternatives, although it has nothing to do with the condition of the same name introduced by Kenneth Arrow in the theory of social choice.

[20] For example, in the example considered, it is rational to agree to ⟨50,50⟩ according to Gauthier. In a choice situation in which only ⟨0,0⟩, ⟨100,30⟩, and ⟨50,50⟩ are feasible – but not ⟨0,100⟩ – then it is rational according to Gauthier to agree to ⟨100,30⟩ (and not ⟨50,50⟩). This is because person two's maximum gain drops (from 100) to 50, and consequently, the maximum relative concession of ⟨100,30⟩ changes from 0.7 (= 70/100) to 0.4 (= 20/50), whereas that of ⟨50,50⟩ remains at 0.5 (= 50/100). This violates Condition Alpha, which requires that if ⟨50,50⟩ is the rational choice in the first case, it must also be the rational choice in the second case (in which ⟨0,100⟩ is not feasible).

to agree to. There remains, however, a significant problem. It is one thing to agree to cooperate (e.g., to help each other paint our houses), quite another to *comply* with that agreement (e.g., to help you paint your house after you have already helped me paint mine). Although, in general, it may be in one's self-interest to comply with agreements, it seems that, at least sometimes, it is in one's interest not to comply. Why should anyone *comply* with the terms of a rational agreement? More specifically, does rationality *always* require that we comply with rational agreements?

Gauthier thinks so. He argues that under certain broadly characterized conditions, rationality requires that fully informed, rational agents adopt a policy ("choice disposition") of complying with the terms of rational agreements. More specifically, he argues that if there are enough agents disposed to comply with rational agreements, and if our characters are sufficiently translucent (in that other people can have a fairly good idea what we are really like, and likely to do), it is in our self-interest to choose to adopt the policy of *constrained maximization* (maximizing our own utility subject to the constraint that we keep rational agreements with others who are disposed to keep rational agreements) rather than dispose ourselves to be *straightforward maximizers* (maximizing our own utility, even when it involves breaking a rational agreement). For if our characters are sufficiently translucent, and we have not adopted a policy of complying with rational agreements, we will be excluded from beneficial cooperative arrangements, because others will not trust us (they will see that we won't keep the agreements we make). Thus, rationality requires us to adopt a policy of complying.

Furthermore, a choice is rationally permissible, Gauthier claims, if and only if it conforms with a policy that it is rational to adopt. Consequently, under the specified conditions, a choice is rationally permissible only if it complies with rational agreements.

There are thus two main claims in Gauthier's argument. The first is that the policy (choice disposition) of constrained maximization is the most advantageous policy to adopt. Here the issue is whether the policy of straightforward maximization, *or some other policy*, is the most advantageous. The second main claim is that the rational permissibility of a choice is determined by whether it conforms with a policy that it would be rational to adopt. This contradicts the received view that the rationality of a choice is determined by whether *it* (as opposed to some policy to which it conforms) maximizes the agent's utility.

This is an extremely important argument, for it purports to show that there is a rational solution to the problem of compliance. If successful, it follows that no enforcement mechanism (that imposes sanctions on those that do not comply) is needed to ensure compliance among fully rational agents. All we need to do, Gauthier argues, is to properly understand the dictates of rationality.

Gauthier's argument that rationality requires that we comply with our rational agreements is especially important in the context of his contractarian moral theory. According to contractarianism, an action is morally permissible if and only if it conforms to the code of conduct to which it would be rational for the members of society to agree. Thus, if rationality requires compliance with rational agreements, and moral constraints are the object of a rational agreement, then rationality requires that one comply with moral constraints. Gauthier thus has an answer to the age-old question, "Why should one be moral?"

Conclusion

In broad outline, then, Gauthier has three main projects: defending a specific contractarian theory of morality; defending a specific instrumental theory of rationality; and defending the claim that under a broad range of circumstances, rationality requires one to act morally. These are ambitious projects well worth examining carefully. For even if unsuccessful, collectively they represent one of the most carefully articulated and rigorously defended positions on the connection between rationality and morality.[21]

[21] This essay has been drawn largely from my "Gauthier on Rationality and Morality," *Eidos* 5 (1986): 79–95. I have benefitted from the comments of the following people: David Braybrooke, Peter Danielson, Morry Lipson, Chris Morris, Jan Narveson, Geoff Sayre-McCord, and an anonymous referee for Cambridge University Press.

Part I

Gauthier's contractarian moral theory

Overview of the essays

In his essay, Gauthier elaborates on his view that moral judgments require a deliberative justification (i.e., one based on the norms of rational choice). Moral judgments require a deliberative justification, because we no longer believe that there are objective moral values, and because, in any case, deliberative justification is more basic (better reflects our deep sense of self). Morality is not, however, to be rejected. Rather, a place for moral constraint can be found within (but not outside) the framework of deliberative justification. Moral constraints can be identified with those constraints on conduct to which the members of society would agree. Given that deliberative rationality requires that one keep one's rational agreements, it also requires that one comply with the constraints of a contractarian morality. Gauthier also elaborates on his defense of the relevance of hypothetical (rather than actual) agreement in a presocial (rather than a social) context.

In her essay, Jean Hampton contrasts the Hobbesian and the Kantian approaches to contractarian theory. She lays out the core of Hobbes's theory, and then assesses the success of Gauthier's theory as a modern day Hobbesian moral theory. She criticizes Gauthier's use of the Lockean Proviso in the specification of the initial bargaining position. The noncooperative outcome – not Gauthier's noncoercive (i.e., constrained by the Lockean Proviso) noncooperative outcome – is, she argues, not only more Hobbesian, but also the appropriate baseline for rational agreement. She then argues that Gauthier's recognition of the role of socialization in forming individuals (e.g., their beliefs, desires, and capacities) leads him to abandon the radical individualism of Hobbes. It leads him in particular, she suggests, to view people as intrinsically valuable – not

merely instrumentally valuable as on Hobbes's view. And this leads, she claims, to a more Kantian – and more promising – contractarian theory.

David Braybrooke examines some of the implications of Gauthier's theory for moral progress in the selection of social rules. He argues that the implications are mixed. Unlike Rawls' contractarian theory, Gauthier's theory cannot be relied upon to give much protection to civil liberties. Furthermore, Gauthier's assumption that the agents are fully informed precludes a Mill-like defense of liberty as instrumentally valuable for the accommodation of future discoveries about one's preferences. On the other hand, under appropriate circumstances (e.g., imperfect competition in the market, or greater human productivity under a planning mechanism), Gauthier's theory would justify a planned (as opposed to a market) economy, and even justify a revolution to bring it about. But it could just as well justify the abandonment of the week and poor by the strong and rich on the grounds that the former no longer offer benefits in cooperation. Finally, although Braybrooke allows that Gauthier's theory falls in with humane feeling in encouraging moral progress, he argues that Gauthier's agents may not in fact care about advances in legal justice, and may care unduly about other people's private lives. In both respects, Gauthier's contract might revoke recent moral gains.

Peter Vallentyne argues that Gauthier's assumption that the parties to the agreement are mutually unconcerned (take no interest in the interests of others) is incompatible with Gauthier's project of reducing morality to rationality. For if morality is to be *reduced* to rationality, people's actual preference must be taken into account. Since people's actual preferences are not mutually unconcerned, an agreement that would be rational if the parties were mutually unconcerned need not be rational given people's actual preferences.

Chris Morris explores the manner in which contractarian theory in general, and Gauthier's theory in particular, determines what has moral standing, that is, what sorts of things are owed moral consideration. It might seem that only the rational agents participating in the social agreement have moral standing on the contractarian view. Given that, on Gauthier's view, children, the significantly infirm, and animals are not participants in the social agreement (because their cooperation offers no benefit to others), it thus seems they do not have moral standing. Morris argues that this is not so. He distinguishes between primary moral standing (which does not depend on anyone's caring about the entity's interests) and secondary moral standing (which does), and then shows that there is room in Gauthier's theory for infants, the infirm, and animals, for example, to have secondary moral standing, provided that the desires of the parties to the agreement are appropriately other-regarding.

2. Why contractarianism?*

David Gauthier

I

As the will to truth thus gains self-consciousness – there can be no doubt of that – morality will gradually *perish* now: this is the great spectacle in a hundred acts reserved for the next two centuries in Europe – the most terrible, most questionable, and perhaps also the most hopeful of all spectacles.

– Nietzsche[1]

Morality faces a foundational crisis. Contractarianism offers the only plausible resolution of this crisis. These two propositions state my theme. What follows is elaboration.

Nietzsche may have been the first, but he has not been alone, in recognizing the crisis to which I refer. Consider these recent statements. "The hypothesis which I wish to advance is that in the actual world which we inhabit the language of morality is in . . . [a] state of grave disorder . . . we have – very largely, if not entirely – lost our comprehension, both theoretical and practical, of morality" (Alasdair MacIntyre).[2] "The resources of most modern moral philosophy are not well adjusted to the modern world" (Bernard Williams).[3] "There are no ob-

*Two paragraphs of Section II and most of Section IV are taken from "Morality, Rational Choice, and Semantic Representation – A Reply to My Critics," in E. F. Paul, F. D. Miller, Jr., and J. Paul (eds.), *The New Social Contract: Essays on Gauthier* (Oxford: Blackwell, 1988), pp. 173–4, 179–180, 184–5, 188–9 (this volume appears also as *Social Philosophy and Policy* 5 [1988], same pagination). I am grateful to Annette Baier, Paul Hurley, and Geoffrey Sayre-McCord for comments on an earlier draft. I am also grateful to discussants at Western Washington University, the University of Arkansas, the University of California at Santa Cruz, and the University of East Anglia for comments on a related talk.

[1] *On the Genealogy of Morals*, trans. by Walter Kaufmann and R. J. Hollingdale (New York: Random House, 1967), third essay, sec. 27, p. 161.

[2] *After Virtue* (Notre Dame, IN: University of Notre Dame Press, 1981), p. 2.

[3] *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press, 1985), p. 197.

jective values. . . . [But] the main tradition of European moral philosophy includes the contrary claim" (J. L. Mackie).[4] "Moral hypotheses do not help explain why people observe what they observe. So ethics is problematic and nihilism must be taken seriously. . . . An extreme version of nihilism holds that morality is simply an illusion. . . . In this version, we should abandon morality, just as an atheist abandons religion after he has decided that religious facts cannot help explain observations" (Gilbert Harman).[5]

I choose these statements to point to features of the crisis that morality faces. They suggest that moral language fits a world view that we have abandoned – a view of the world as purposively ordered. Without this view, we no longer truly understand the moral claims we continue to make. They suggest that there is a lack of fit between what morality presupposes – objective values that help explain our behavior, and the psychological states – desires and beliefs – that, given our present world view, actually provide the best explanation. This lack of fit threatens to undermine the very idea of a morality as more than an anthropological curiosity. But how could this be? How could morality *perish*?

II

To proceed, I must offer a minimal characterization of the morality that faces a foundational crisis. And this is the morality of justified constraint. From the standpoint of the agent, moral considerations present themselves as constraining his choices and actions, in ways independent of his desires, aims, and interests. Later, I shall add to this characterization, but for the moment it will suffice. For it reveals clearly what is in question – the ground of constraint. This ground seems absent from our present world view. And so we ask, what reason can a person have for recognizing and accepting a constraint that is independent of his desires and interests? He may agree that such a constraint would be *morally* justified; he would have a reason for accepting it *if* he had a reason for accepting morality. But what justifies paying attention to morality, rather than dismissing it as an appendage of outworn beliefs? We ask, and seem to find no answer. But before proceeding, we should consider three objections.

The first is to query the idea of constraint. Why should morality be seen as constraining our choices and actions? Why should we not rather say that the moral person chooses most freely, because she chooses in the light of a true conception of herself, rather than in the light of the false conceptions that so often predominate? Why should we not link

[4] *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin, 1977), pp. 15, 30.
[5] *The Nature of Morality* (New York: Oxford University Press, 1977), p. 11.

morality with self-understanding? Plato and Hume might be enlisted to support this view, but Hume would be at best a partial ally, for his representation of "virtue in all her genuine and most engaging charms, . . . talk[ing] not of useless austerities and rigors, suffering and self-denial," but rather making "her votaries . . . , during every instant of their existence, if possible, cheerful and happy," is rather overcast by his admission that "in the case of justice, . . . a man, taking things in a certain light, may often seem to be a loser by his integrity."[6] Plato, to be sure, goes further, insisting that only the just man has a healthy soul, but heroic as Socrates' defense of justice may be, we are all too apt to judge that Glaucon and Adeimantus have been charmed rather than reasoned into agreement, and that the unjust man has not been shown necessarily to be the loser.[7] I do not, in any event, intend to pursue this direction of thought. Morality, as we, heirs to the Christian and Kantian traditions, conceive it, constrains the pursuits to which even our reflective desires would lead us. And this is not simply or entirely a constraint on self-interest; the affections that morality curbs include the social ones of favoritism and partiality, to say nothing of cruelty.

The second objection to the view that moral constraint is insufficiently grounded is to query the claim that it operates independently of, rather than through, our desires, interests, and affections. Morality, some may say, concerns the well-being of all persons, or perhaps of all sentient creatures.[8] And one may then argue, either with Hume, that morality arises in and from our sympathetic identification with our fellows, or that it lies directly in well-being, and that our affections tend to be disposed favorably toward it. But, of course, not all of our affections. And so our sympathetic feelings come into characteristic opposition to other feelings, in relation to which they function as a constraint.

This is a very crude characterization, but it will suffice for the present argument. This view grants that morality, as we understand it, is without purely *rational* foundations, but reminds us that we are not therefore unconcerned about the well-being of our fellows. Morality is founded on the widespread, sympathetic, other-directed concerns that most of us have, and these concerns do curb self-interest, and also the favoritism and partiality with which we often treat others. Nevertheless, if morality depends for its practical relevance and motivational efficacity entirely on our sympathetic feelings, it has no title to the prescriptive grip with which it has been invested in the Christian and Kantian views to which I have referred, and which indeed Glaucon and Adeimantus demanded

[6] David Hume, *An Equiry Concerning the Principles of Morals*, 1751, sec. IX, pt. II.
[7] See Plato, *Republic*, esp. books II and IV.
[8] Some would extend morality to the nonsentient, but sympathetic as I am to the rights of trolley cars and steam locomotives, I propose to leave this view quite out of consideration.

that Socrates defend to them in the case of justice. For to be reminded that some of the time we do care about our fellows and are willing to curb other desires in order to exhibit that care tells us nothing that can guide us in those cases in which, on the face of it, we do not care, or do not care enough – nothing that will defend the demands that morality makes on us in the hard cases. That not all situations in which concern for others combats self-concern are hard cases is true, but morality, as we ordinarily understand it, speaks to the hard cases, whereas its Humean or naturalistic replacement does not.

These remarks apply to the most sustained recent positive attempt to create a moral theory – that of John Rawls. For the attempt to describe our moral capacity, or more particularly, for Rawls, our sense of justice, in terms of principles, plausible in the light of our more general psychological theory, and coherent with "our considered judgments in reflective equilibrium,"[9] will not yield any answer to why, in those cases in which we have no, or insufficient, interest in being just, we should nevertheless follow the principles. John Harsanyi, whose moral theory is in some respects a utilitarian variant of Rawls' contractarian construction, recognizes this explicitly: "All we can prove by rational arguments is that anybody who wants to serve our common human interests in a rational manner must obey these commands."[10] But although morality may offer itself in the service of our common human interests, it does not offer itself only to those who want to serve them.

Morality is a constraint that, as Kant recognized, must not be supposed to depend solely on our feelings. And so we may not appeal to feelings to answer the question of its foundation. But the third objection is to dismiss this question directly, rejecting the very idea of a foundational crisis. Nothing justifies morality, for morality needs no justification. We find ourselves, in morality as elsewhere, in mediis rebus. We make, accept and reject, justify and criticize moral judgments. The concern of moral theory is to systematize that practice, and so to give us a deeper understanding of what moral justification is. But there are no extramoral foundations for moral justification, any more than there are extraepistemic foundations for epistemic judgments. In morals as in science, foundationalism is a bankrupt project.

Fortunately, I do not have to defend *normative* foundationalism. One problem with accepting moral justification as part of our ongoing practice is that, as I have suggested, we no longer accept the world view on which it depends. But perhaps a more immediately pressing problem is that we have, ready to hand, an alternative mode for justifying our

[9] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), p. 51.
[10] John C. Harsanyi, "Morality and the Theory of Rational Behaviour," in *Utilitarianism and Beyond*, edited by Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982), p. 62.

choices and actions. In its more austere and, in my view, more defensible form, this is to show that choices and actions maximize the agent's expected utility, where utility is a measure of considered preference. In its less austere version, this is to show that choices and actions satisfy, not a subjectively defined requirement such as utility, but meet the agent's objective interests. Since I do not believe that we have objective interests, I shall ignore this latter. But it will not matter. For the idea is clear; we have a mode of justification that does not require the introduction of moral considerations.[11]

Let me call this alternative nonmoral mode of justification, neutrally, deliberative justification. Now moral and deliberative justification are directed at the same objects – our choices and actions. What if they conflict? And what do we say to the person who offers a deliberative justification of his choices and actions and refuses to offer any other? We can say, of course, that his behavior lacks *moral* justification, but this seems to lack any hold, unless he chooses to enter the moral framework. And such entry, he may insist, lacks any deliberative justification, at least for him.

If morality perishes, the justificatory enterprise, in relation to choice and action, does not perish with it. Rather, one mode of justification perishes, a mode that, it may seem, now hangs unsupported. But not only unsupported, for it is difficult to deny that deliberative justification is more clearly basic, that it cannot be avoided insofar as we are rational agents, so that if moral justification conflicts with it, morality seems not only unsupported but opposed by what is rationally more fundamental.

Deliberative justification relates to our deep sense of self. What distinguishes human beings from other animals, and provides the basis for rationality, is the capacity for semantic representation. You can, as your dog on the whole cannot, represent a state of affairs to yourself, and consider in particular whether or not it is the case, and whether or not you would want it to be the case. You can represent to yourself the contents of your beliefs, and your desires or preferences. But in representing them, you bring them into relation with one another. You represent to yourself that the Blue Jays will win the World Series, and that a National League team will win the World Series, and that the Blue Jays are not a National League team. And in recognizing a conflict among those beliefs, you find rationality thrust upon you. Note that the first two beliefs could be replaced by preferences, with the same effect.

Since in representing our preferences we become aware of conflict among them, the step from representation to choice becomes compli-

[11] To be sure, if we think of morality as expressed in certain of our affections and/or interests, it will incorporate moral considerations to the extent that they actually are present in our preferences. But this would be to embrace the naturalism that I have put to one side as inadequate.

cated. We must, somehow, bring our conflicting desires and preferences into some sort of coherence. And there is only one plausible candidate for a principle of coherence – a maximizing principle. We order our preferences, in relation to decision and action, so that we may choose in a way that maximizes our expectation of preference fulfillment. And in so doing, we show ourselves to be rational agents, engaged in deliberation and deliberative justification. There is simply nothing else for practical rationality to be.

The foundational crisis of morality thus cannot be avoided by pointing to the existence of a practice of justification within the moral framework, and denying that any extramoral foundation is relevant. For an extramoral mode of justification is already present, existing not side by side with moral justification, but in a manner tied to the way in which we unify our beliefs and preferences and so acquire our deep sense of self. We need not suppose that this deliberative justification is itself to be understood foundationally. All that we need suppose is that moral justification does not plausibly survive conflict with it.

III

In explaining why we may not dismiss the idea of a foundational crisis in morality as resulting from a misplaced appeal to a philosophically discredited or suspect idea of foundationalism, I have begun to expose the character and dimensions of the crisis. I have claimed that morality faces an alternative, conflicting, deeper mode of justification, related to our deep sense of self, that applies to the entire realm of choice and action, and that evaluates each *action* in terms of the reflectively held concerns of its *agent*. The relevance of the agent's concerns to practical justification does not seem to me in doubt. The relevance of anything else, except insofar as it bears on the agent's concerns, does seem to me very much in doubt. If the agent's reflectively endorsed concerns, his preferences, desires, and aims, are, with his considered beliefs, constitutive of his self-conception, then I can see no remotely plausible way of arguing from their relevance to that of anything else that is not similarly related to his sense of self. And, indeed, I can see no way of introducing anything as relevant to practical justification except through the agent's self-conception. My assertion of this practical individualism is not a conclusive argument, but the burden of proof is surely on those who would maintain a contrary position. Let them provide the arguments – if they can.

Deliberative justification does not refute morality. Indeed, it does not offer morality the courtesy of a refutation. It ignores morality, and seemingly replaces it. It preempts the arena of justification, apparently leaving morality no room to gain purchase. Let me offer a controversial comparison. Religion faces – indeed, has faced – a comparable foundational crisis. Religion demands the worship of a divine being who purposively orders the universe. But it has confronted an alternative mode of explanation. Although the emergence of a cosmological theory based on efficient, rather than teleological, causation provided warning of what was to come, the supplanting of teleology in biology by the success of evolutionary theory in providing a mode of explanation that accounted in efficient-causal terms for the *appearance* of a purposive order among living beings, may seem to toll the death knell for religion as an intellectually respectable enterprise. But evolutionary biology and, more generally, modern science do not refute religion. Rather they ignore it, replacing its explanations by ontologically simpler ones. Religion, understood as affirming the justifiable worship of a divine being, may be unable to survive its foundational crisis. Can morality, understood as affirming justifiable constraints on choice independent of the agent's concerns, survive?

There would seem to be three ways for morality to escape religion's apparent fate. One would be to find, for moral facts or moral properties, an explanatory role that would entrench them prior to any consideration of justification.[12] One could then argue that any mode of justification that ignored moral considerations would be ontologically defective. I mention this possibility only to put it to one side. No doubt there are persons who accept moral constraints on their choices and actions, and it would not be possible to explain those choices and actions were we to ignore this. But our explanation of their behavior need not commit us to their view. Here the comparison with religion should be straightforward and uncontroversial. We could not explain many of the practices of the religious without reference to their beliefs. But to characterize what a religious person is doing as, say, an act of worship, does not commit us to supposing that an object of worship actually exists, though it does commit us to supposing that she believes such an object to exist. Similarly, to characterize what a moral agent is doing as, say, fulfilling a duty does not commit us to supposing that there are any duties, though it does commit us to supposing that he believes that there are duties. The skeptic who accepts neither can treat the apparent role of morality in explanation as similar to that of religion. Of course, I do not consider that the parallel can be ultimately sustained, since I agree with the religious skeptic but not with the moral skeptic. But to establish an explanatory role for morality, one must first demonstrate its justificatory credentials. One may not assume that it has a prior explanatory role.

The second way would be to reinterpret the idea of justification, show-

[12] This would meet the challenge to morality found in my previous quotation from Gilbert Harman.

ing that, more fully understood, deliberative justification is incomplete, and must be supplemented in a way that makes room for morality. There is a long tradition in moral philosophy, deriving primarily from Kant, that is committed to this enterprise. This is not the occasion to embark on a critique of what, in the hope again of achieving a neutral characterization, I shall call universalistic justification. But critique may be out of place. The success of deliberative justification may suffice. For theoretical claims about its incompleteness seem to fail before the simple practical recognition that it works. Of course, on the face of it, deliberative justification does not work to provide a place for morality. But to suppose that it must, if it is to be fully adequate or complete as a mode of justification, would be to assume what is in question, whether moral justification is defensible.

If, independent of one's actual desires, and aims, there were objective values, and if, independent of one's actual purposes, one were part of an objectively purposive order, then we might have reason to insist on the inadequacy of the deliberative framework. An objectively purposive order would introduce considerations relevant to practical justification that did not depend on the agent's self-conception. But the supplanting of teleology in our physical and biological explanations closes this possibility, as it closes the possibility of religious explanation.

I turn then to the third way of resolving morality's foundational crisis. The first step is to embrace deliberative justification, and recognize that morality's place must be found within, and not outside, its framework. Now this will immediately raise two problems. First of all, it will seem that the attempt to establish any constraint on choice and action, within the framework of a deliberation that aims at the maximal fulfillment of the agent's considered preferences, must prove impossible. But even if this be doubted, it will seem that the attempt to establish a constraint *independent of the agent's preferences*, within such a framework, verges on lunacy. Nevertheless, this is precisely the task accepted by my third way. And, unlike its predecessors, I believe that it can be successful; indeed, I believe that my recent book, *Morals by Agreement*, shows how it can succeed.[13]

I shall not rehearse at length an argument that is now familiar to at least some readers, and, in any event, can be found in that book. But let me sketch briefly those features of deliberative rationality that enable it to constrain maximizing choice. The key idea is that in many situations, if each person chooses what, given the choices of the others, would maximize her expected utility, then the outcome will be mutually disadvantageous in comparison with some alternative – everyone could do

[13] See David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986), especially chaps. V and VI.

better.[14] Equilibrium, which obtains when each person's action is a best response to the others' actions, is incompatible with (Pareto-)optimality, which obtains when no one could do better without someone else doing worse. Given the ubiquity of such situations, each person can see the benefit, to herself, of participating with her fellows in practices requiring each to refrain from the direct endeavor to maximize her own utility, when such mutual restraint is mutually advantageous. No one, of course, can have reason to accept any unilateral constraint on her maximizing behavior; each benefits from, and only from, the constraint accepted by her fellows. But if one benefits more from a constraint on others than one loses by being constrained oneself, one may have reason to accept a practice requiring everyone, including oneself, to exhibit such a constraint. We may represent such a practice as capable of gaining unanimous agreement among rational persons who were choosing the terms on which they would interact with each other. And this agreement is the basis of morality.

Consider a simple example of a moral practice that would command rational agreement. Suppose each of us were to assist her fellows only when either she could expect to benefit herself from giving assistance, or she took a direct interest in their well-being. Then, in many situations, persons would not give assistance to others, even though the benefit to the recipient would greatly exceed the cost to the giver, because there would be no provision for the giver to share in the benefit. Everyone would then expect to do better were each to give assistance to her fellows, regardless of her own benefit or interest, whenever the cost of assisting was low and the benefit of receiving assistance considerable. Each would thereby accept a constraint on the direct pursuit of her own concerns, not unilaterally, but given a like acceptance by others. Reflection leads us to recognize that those who belong to groups whose members adhere to such a practice of mutual assistance enjoy benefits in interaction that are denied to others. We may then represent such a practice as rationally acceptable to everyone.

This rationale for agreed constraint makes no reference to the content of anyone's preferences. The argument depends simply on the *structure* of interaction, on the way in which each person's endeavor to fulfill her own preferences affects the fulfillment of everyone else. Thus, each person's reason to accept a mutually constraining practice is independent of her particular desires, aims and interests, although not, of course, of the fact that she has such concerns. The idea of a purely rational agent,

[14] The now-classic example of this type of situation is the Prisoner's Dilemma; see *Morals by Agreement*, pp. 79–80. More generally, such situations may be said, in economists' parlance, to exhibit market failure. See, for example, "Market Contractarianism" in Jules Coleman, *Markets, Morals, and the Law* (Cambridge: Cambridge University Press, 1988), chap. 10.

moved to act by reason alone, is not, I think, an intelligible one. Morality is not to be understood as a constraint arising from reason alone on the fulfillment of nonrational preferences. Rather, a rational agent is one who acts to achieve the maximal fulfillment of her preferences, and morality is a constraint on the manner in which she acts, arising from the effects of interaction with other agents.

Hobbes's Foole now makes his familiar entry onto the scene, to insist that however rational it may be for a person to agree with her fellows to practices that hold out the promise of mutual advantage, yet it is rational to follow such practices only when so doing directly conduces to her maximal preference fulfillment.[15] But then such practices impose no real constraint. The effect of agreeing to or accepting them can only be to change the expected payoffs of her possible choices, making it rational for her to choose what in the absence of the practice would not be utility maximizing. The practices would offer only true prudence, not true morality.

The Foole is guilty of a twofold error. First, he fails to understand that real acceptance of such moral practices as assisting one's fellows, or keeping one's promises, or telling the truth is possible only among those who are disposed to comply with them. If my disposition to comply extends only so far as my interests or concerns at the time of performance, then you will be the real fool if you interact with me in ways that demand a more rigorous compliance. If, for example, it is rational to keep promises only when so doing is directly utility maximizing, then among persons whose rationality is common knowledge, only promises that require such limited compliance will be made. And opportunities for mutual advantage will be thereby forgone.

Consider this example of the way in which promises facilitate mutual benefit. Jones and Smith have adjacent farms. Although neighbors, and not hostile, they are also not friends, so that neither gets satisfaction from assisting the other. Nevertheless, they recognize that, if they harvest their crops together, each does better than if each harvests alone. Next week, Jones's crop will be ready for harvesting; a fortnight hence, Smith's crop will be ready. The harvest in, Jones is retiring, selling his farm, and moving to Florida, where he is unlikely to encounter Smith or other members of their community. Jones would like to promise Smith that, if Smith helps him harvest next week, he will help Smith harvest in a fortnight. But Jones and Smith both know that in a fortnight, helping Smith would be a pure cost to Jones. Even if Smith helps him, he has nothing to gain by returning the assistance, since neither care for Smith nor, in the circumstances, concern for his own reputation, moves him. Hence, if Jones and Smith know that Jones acts straightforwardly to

[15] See Hobbes, *Leviathan*, London, 1651, chap. 15.

maximize the fulfillment of his preferences, they know that he will not help Smith. Smith, therefore, will not help Jones even if Jones pretends to promise assistance in return. Nevertheless, Jones would do better could he make and keep such a promise – and so would Smith.

The Foole's second error, following on his first, should be clear; he fails to recognize that in plausible circumstances, persons who are genuinely disposed to a more rigorous compliance with moral practices than would follow from their interests at the time of performance can expect to do better than those who are not so disposed. For the former, constrained maximizers as I call them, will be welcome partners in mutually advantageous cooperation, in which each relies on the voluntary adherence of the others, from which the latter, straightforward maximizers, will be excluded. Constrained maximizers may thus expect more favorable opportunities than their fellows. Although in assisting their fellows, keeping their promises, and complying with other moral practices, they forgo preference fulfillment that they might obtain, yet they do better overall than those who always maximize expected utility, because of their superior opportunities.

In identifying morality with those constraints that would obtain agreement among rational persons who were choosing their terms of interaction, I am engaged in rational reconstruction. I do not suppose that we have actually agreed to existent moral practices and principles. Nor do I suppose that all existent moral practices would secure our agreement, were the question to be raised. Not all existent moral practices need be justifiable – need be ones with which we ought willingly to comply. Indeed, I do not even suppose that the practices with which we ought willingly to comply need be those that would secure our present agreement. I suppose that justifiable moral practices are those that would secure our agreement ex ante, in an appropriate premoral situation. They are those to which we should have agreed as constituting the terms of our future interaction, had we been, per impossible, in a position to decide those terms. Hypothetical agreement thus provides a test of the justifiability of our existent moral practices.

IV

Many questions could be raised about this account, but here I want to consider only one. I have claimed that moral practices are rational, even though they constrain each person's attempt to maximize her own utility, insofar as they would be the objects of unanimous ex ante agreement. But to refute the Foole, I must defend not only the rationality of agreement, but also that of compliance, and the defense of compliance threatens to preempt the case for agreement, so that my title should be "Why Constraint?" and not "Why Contractarianism?" It is rational to dispose

oneself to accept certain constraints on direct maximization in choosing and acting, if and only if so disposing oneself maximizes one's expected utility. What then is the relevance of agreement, and especially of hypothetical agreement? Why should it be rational to dispose oneself to accept only those constraints that would be the object of mutual agreement in an appropriate premoral situation, rather than those constraints that are found in our existent moral practices? Surely it is acceptance of the latter that makes a person welcome in interaction with his fellows. For compliance with existing morality will be what they expect, and take into account in choosing partners with whom to cooperate.

I began with a challenge to morality – how can it be rational for us to accept its constraints? It may now seem that what I have shown is that it is indeed rational for us to accept constraints, but to accept them whether or not they might be plausibly considered moral. Morality, it may seem, has nothing to do with my argument; what I have shown is that it is rational to be disposed to comply with whatever constraints are generally accepted and expected, regardless of their nature. But this is not my view.

To show the relevance of agreement to the justification of constraints, let us assume an ongoing society in which individuals more or less acknowledge and comply with a given set of practices that constrain their choices in relation to what they would be did they take only their desires, aims, and interests directly into account. Suppose that a disposition to conform to these existing practices is prima facie advantageous, since persons who are not so disposed may expect to be excluded from desirable opportunities by their fellows. However, the practices themselves have, or at least need have, no basis in agreement. And they need satisfy no intuitive standard of fairness or impartiality, characteristics that we may suppose relevant to the identification of the practices with those of a genuine morality. Although we may speak of the practices as constituting the morality of the society in question, we need not consider them morally justified or acceptable. They are simply practices constraining individual behavior in a way that each finds rational to accept.

Suppose now that our persons, as rational maximizers of individual utility, come to reflect on the practices constituting their morality. They will, of course, assess the practices in relation to their own utility, but with the awareness that their fellows will be doing the same. And one question that must arise is: Why these practices? For they will recognize that the set of actual moral practices is not the only possible set of constraining practices that would yield mutually advantageous, optimal outcomes. They will recognize the possibility of alternative moral orders. At this point it will not be enough to say that, as a matter of fact, each person can expect to benefit from a disposition to comply with existing

practices. For persons will also ask themselves: Can I benefit more, not from simply abandoning any morality, and recognizing no constraint, but from a partial rejection of existing constraints in favor of an alternative set? Once this question is asked, the situation is transformed; the existing moral order must be assessed, not only against simple noncompliance, but also against what we may call alternative compliance.

To make this assessment, each will compare her prospects under the existing practices with those she would anticipate from a set that, in the existing circumstances, she would expect to result from bargaining with her fellows. If her prospects would be improved by such negotiation, then she will have a real, although not necessarily sufficient, incentive to demand a change in the established moral order. More generally, if there are persons whose prospects would be improved by renegotiation, then the existing order will be recognizably unstable. No doubt those whose prospects would be worsened by renegotiation will have a clear incentive to resist, to appeal to the status quo. But their appeal will be a weak one, especially among persons who are not taken in by spurious ideological considerations, but focus on individual utility maximization. Thus, although in the real world, we begin with an existing set of moral practices as constraints on our maximizing behavior, yet we are led by reflection to the idea of an amended set that would obtain the agreement of everyone, and this amended set has, and will be recognized to have, a stability lacking in existing morality.

The reflective capacity of rational agents leads them from the given to the agreed, from existing practices and principles requiring constraint to those that would receive each person's assent. The same reflective capacity, I claim, leads from those practices that would be agreed to, in existing social circumstances, to those that would receive ex ante agreement, premoral and presocial. As the status quo proves unstable when it comes into conflict with what would be agreed to, so what would be agreed to proves unstable when it comes into conflict with what would have been agreed to in an appropriate presocial context. For as existing practices must seem arbitrary insofar as they do not correspond to what a rational person would agree to, so what such a person would agree to in existing circumstances must seem arbitrary in relation to what she would accept in a presocial condition.

What a rational person would agree to in existing circumstances depends in large part on her negotiating position vis-à-vis her fellows. But her negotiating position is significantly affected by the existing social institutions, and so by the currently accepted moral practices embodied in those institutions. Thus, although agreement may well yield practices differing from those embodied in existing social institutions, yet it will be influenced by those practices, which are not themselves the product of rational agreement. And this must call the rationality of the agreed

practices into question. The arbitrariness of existing practices must infect any agreement whose terms are significantly affected by them. Although rational agreement is in itself a source of stability, yet this stability is undermined by the arbitrariness of the circumstances in which it takes place. To escape this arbitrariness, rational persons will revert from actual to hypothetical agreement, considering what practices they would have agreed to from an initial position not structured by existing institutions and the practices they embody.

The content of a hypothetical agreement is determined by an appeal to the equal rationality of persons. Rational persons will voluntarily accept an agreement only insofar as they perceive it to be equally advantageous to each. To be sure, each would be happy to accept an agreement more advantageous to herself than to her fellows, but since no one will accept an agreement perceived to be less advantageous, agents whose rationality is a matter of common knowledge will recognize the futility of aiming at or holding out for more, and minimize their bargaining costs by coordinating at the point of equal advantage. Now the extent of advantage is determined in a twofold way. First, there is advantage internal to an agreement. In this respect, the expectation of equal advantage is assured by procedural fairness. The step from existing moral practices to those resulting from actual agreement takes rational persons to a procedurally fair situation, in which each perceives the agreed practices to be ones that it is equally rational for all to accept, given the circumstances in which agreement is reached. But those circumstances themselves may be called into question insofar as they are perceived to be arbitrary – the result, in part, of compliance with constraining practices that do not themselves ensure the expectation of equal advantage, and so do not reflect the equal rationality of the complying parties. To neutralize this arbitrary element, moral practices to be fully acceptable must be conceived as constituting a possible outcome of a hypothetical agreement under circumstances that are unaffected by social institutions that themselves lack full acceptability. Equal rationality demands consideration of external circumstances as well as internal procedures.

But what is the practical import of this argument? It would be absurd to claim that mere acquaintance with it, or even acceptance of it, will lead to the replacement of existing moral practices by those that would secure presocial agreement. It would be irrational for anyone to give up the benefits of the existing moral order simply because he comes to realize that it affords him more than he could expect from pure rational agreement with his fellows. And it would be irrational for anyone to accept a long-term utility loss by refusing to comply with the existing moral order, simply because she comes to realize that such compliance affords her less than she could expect from pure rational agreement.

Nevertheless, these realizations do transform, or perhaps bring to the surface, the character of the relationships between persons that are maintained by the existing constraints, so that some of these relationships come to be recognized as coercive. These realizations constitute the elimination of false consciousness, and they result from a process of rational reflection that brings persons into what, in my theory, is the parallel of Jürgen Habermas's ideal speech situation.[16] Without an argument to defend themselves in open dialogue with their fellows, those who are more than equally advantaged can hope to maintain their privileged position only if they can coerce their fellows into accepting it. And this, of course, may be possible. But coercion is not agreement, and it lacks any inherent stability.

Stability plays a key role in linking compliance to agreement. Aware of the benefits to be gained from constraining practices, rational persons will seek those that invite stable compliance. Now compliance is stable if it arises from agreement among persons each of whom considers both that the terms of agreement are sufficiently favorable to herself that it is rational for her to accept them, and that they are not so favorable to others that it would be rational for them to accept terms less favorable to them and more favorable to herself. An agreement affording equally favorable terms to all thus invites, as no other can, stable compliance.

V

In defending the claim that moral practices, to obtain the stable voluntary compliance of rational individuals, must be the objects of an appropriate hypothetical agreement, I have added to the initial minimal characterization of morality. Not only does morality constrain our choices and actions, but it does so in an impartial way, reflecting the equal rationality of the persons subject to constraint. Although it is no part of my argument to show that the requirements of contractarian morality will satisfy the Rawlsian test of cohering with our considered judgments in reflective equilibrium, yet it would be misleading to treat rationally agreed constraints on direct utility maximization as constituting a morality at all, rather than as replacing morality, were there no fit between their content and our pretheoretical moral views. The fit lies, I suggest, in the impartiality required for hypothetical agreement.

The foundational crisis of morality is thus resolved by exhibiting the rationality of our compliance with mutual, rationally agreed constraints on the pursuit of our desires, aims, and interests. Although bereft of a basis in objective values or an objectively purposive order, and con-

[16] See Raymond Geuss, *The Idea of a Critical Theory: Habermas and the Frankfurt School* (Cambridge: Cambridge University Press, 1981), p. 65ff.

fronted by a more fundamental mode of justification, morality survives by incorporating itself into that mode. Moral considerations have the same status, and the same role in explaining behavior, as the other reasons acknowledged by a rational deliberator. We are left with a unified account of justification, in which an agent's choices and actions are evaluated in relation to his preferences – to the concerns that are constitutive of his sense of self. But since morality binds the agent independently of the particular content of his preferences, it has the prescriptive grip with which the Christian and Kantian views have invested it.

In incorporating morality into deliberative justification, we recognize a new dimension to the agent's self-conception. For morality requires that a person have the capacity to commit himself, to enter into agreement with his fellows secure in the awareness that he can and will carry out his part of the agreement without regard to many of those considerations that normally and justifiably would enter into his future deliberations. And this is more than the capacity to bring one's desires and interests together with one's beliefs into a single coherent whole. Although this latter unifying capacity must extend its attention to past and future, the unification it achieves may itself be restricted to that extended present within which a person judges and decides. But in committing oneself to future action in accordance with one's agreement, one must fix at least a subset of one's desires and beliefs to hold in that future. The self that agrees and the self that complies must be one. "Man himself must first of all have become *calculable, regular, necessary*, even in his own image of himself, if he is to be able to stand security for *his own future*, which is what one who promises does!"[17]

In developing *"the right to make promises,"*[18] we human beings have found a contractarian bulwark against the perishing of morality.

[17] Nietzsche, *On the Genealogy of Morals*, trans. by Walter Kaufmann and R. J. Hollingdale (New York: Random House, 1967), second essay, sec. 1, p. 58.
[18] Ibid., p. 57.

3. Two faces of contractarian thought

Jean Hampton

"... What was I created for, I wonder? Where is my place in the world?"
She mused again.
"Ah! I see," she pursued presently, "that is the question which most old maids are puzzled to solve: other people solve it for them by saying, 'Your place is to do good to others, to be helpful whenever help is wanted.' That is right in some measure, and a very convenient doctrine for the people who hold it; but I perceive that certain sets of human beings are very apt to maintain that other sets should give up their lives to them and their service, and then they requite them by praise: they call them devoted and virtuous. Is this enough? Is it to live? Is there not a terrible hollowness, mockery, want, craving, in that existence which is given away to others, for want of something of your own to bestow it on? I suspect there is. Does virtue lie in abnegation of self? I do not believe it. Undue humility makes tyranny; weak concession creates selfishness. ... Each human being has his share of rights. I suspect it would conduce to the happiness and welfare of all, if each knew his allotment and held to it as tenaciously as a martyr to his creed. Queer thoughts these, that surge in my mind: are they right thoughts? I am not certain."

– Charlotte Brontë, *Shirley*

The quotation that begins this essay is pertinent to its conclusion. But to launch the discussion, I want to quote from nineteenth-century adventurer Mary Kingsley. Dressed in skirts, she traveled alone into the African interior where no white man or woman had ever been, climbing mountains, navigating rivers, fighting animals – and trading her way because, she explained, "when you first appear among people who have never seen anything like you before, they naturally regard you as a devil; but when you want to buy or sell with them, they recognize there is something human and reasonable about you."[1] The idea that the

[1] From Katherine Frank, *A Voyager Out: The Life of Mary Kingsley* (Boston: Houghton Mifflin, 1986), p. 63.

fronted by a more fundamental mode of justification, morality survives by incorporating itself into that mode. Moral considerations have the same status, and the same role in explaining behavior, as the other reasons acknowledged by a rational deliberator. We are left with a unified account of justification, in which an agent's choices and actions are evaluated in relation to his preferences – to the concerns that are constitutive of his sense of self. But since morality binds the agent independently of the particular content of his preferences, it has the prescriptive grip with which the Christian and Kantian views have invested it.

In incorporating morality into deliberative justification, we recognize a new dimension to the agent's self-conception. For morality requires that a person have the capacity to commit himself, to enter into agreement with his fellows secure in the awareness that he can and will carry out his part of the agreement without regard to many of those considerations that normally and justifiably would enter into his future deliberations. And this is more than the capacity to bring one's desires and interests together with one's beliefs into a single coherent whole. Although this latter unifying capacity must extend its attention to past and future, the unification it achieves may itself be restricted to that extended present within which a person judges and decides. But in committing oneself to future action in accordance with one's agreement, one must fix at least a subset of one's desires and beliefs to hold in that future. The self that agrees and the self that complies must be one. "Man himself must first of all have become *calculable, regular, necessary*, even in his own image of himself, if he is to be able to stand security for *his own future*, which is what one who promises does!"[17]

In developing *"the right to make promises,"*[18] we human beings have found a contractarian bulwark against the perishing of morality.

[17] Nietzsche, *On the Genealogy of Morals*, trans. by Walter Kaufmann and R. J. Hollingdale (New York: Random House, 1967), second essay, sec. 1, p. 58.
[18] Ibid., p. 57.

3. Two faces of contractarian thought

Jean Hampton

"... What was I created for, I wonder? Where is my place in the world?"
She mused again.
"Ah! I see," she pursued presently, "that is the question which most old maids are puzzled to solve: other people solve it for them by saying, 'Your place is to do good to others, to be helpful whenever help is wanted.' That is right in some measure, and a very convenient doctrine for the people who hold it; but I perceive that certain sets of human beings are very apt to maintain that other sets should give up their lives to them and their service, and then they requite them by praise: they call them devoted and virtuous. Is this enough? Is it to live? Is there not a terrible hollowness, mockery, want, craving, in that existence which is given away to others, for want of something of your own to bestow it on? I suspect there is. Does virtue lie in abnegation of self? I do not believe it. Undue humility makes tyranny; weak concession creates selfishness. ... Each human being has his share of rights. I suspect it would conduce to the happiness and welfare of all, if each knew his allotment and held to it as tenaciously as a martyr to his creed. Queer thoughts these, that surge in my mind: are they right thoughts? I am not certain."

– Charlotte Brontë, *Shirley*

The quotation that begins this essay is pertinent to its conclusion. But to launch the discussion, I want to quote from nineteenth-century adventurer Mary Kingsley. Dressed in skirts, she traveled alone into the African interior where no white man or woman had ever been, climbing mountains, navigating rivers, fighting animals – and trading her way because, she explained, "when you first appear among people who have never seen anything like you before, they naturally regard you as a devil; but when you want to buy or sell with them, they recognize there is something human and reasonable about you."[1] The idea that the

[1] From Katherine Frank, *A Voyager Out: The Life of Mary Kingsley* (Boston: Houghton Mifflin, 1986), p. 63.

essence of human rationality, and even human morality, is embodied in the notion of a contract is the heart of what is called the "contractarian" approach to moral thinking.

The theory itself goes back hundreds of years.[2] In modern times, Grotius, Suarez, Hobbes, Locke, Rousseau, and Kant used the notion of "what people could agree to" primarily in order to argue for the legitimacy of political institutions with a certain structure and purpose. This explicitly political orientation of the argument is maintained by Robert Nozick in his recent book *Anarchy, State and Utopia*.[3] I call those contractarian theorists who argue for the legitimacy of the state using contract language "state contractarians." Other modern contractarians have not restricted the contractarian form of argument to this one political issue. They are convinced that it can be used to identify and motivate commitment to the best available conception of justice, and perhaps also to other kinds of cooperative behavior that involve constraining one's self-regarding pursuits in ways that benefit the community. I call these sorts of theorists "moral contractarians." John Rawls was one of the first contemporary moral contractarians. Interestingly, Rawls does not even question the state's legitimacy in *A Theory of Justice*,[4] but instead tries to identify the best available conception of justice for the structuring of our political and social institutions by using what he calls a hypothetical contract. Scanlon, Gauthier, Grice, Harman, and Mackie go even further, explicitly endorsing the contractarian argument as a way of understanding not only the duties of justice, but virtually the entire content of morality.

Yet the fact that all of these theorists call their theories "contractarian" is misleading, because that single label masks deep differences among them. The most obvious difference concerns what they take to be the metaethical status of their contractarian moral theories. Consider that two of them – Rawls and Scanlon – have regarded themselves as moral objectivists (although Rawls' thinking has, at least arguably, undergone some change in recent years[5]) and have linked their theorizing in some fashion to Kant, whereas others – for example, Gauthier, Harman, and

[2] See J. W. Gough, *The Social Contract: A Critical Study of Its Development* (Oxford: Clarendon Press, 1936), and the Introduction to my *Hobbes and the Social Contract Tradition* (Cambridge: Cambridge University Press, 1986).

[3] Robert Nozick, *Anarchy, State and Utopia* (New York: Basic Books, 1974).

[4] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971).

[5] In this essay, the Rawls that is discussed is the Rawls who wrote *A Theory of Justice* and not the Rawls of such recent articles as "Justice as Fairness: Political Not Metaphysical" (in *Philosophy and Public Affairs*, Summer 1985). In the book, Rawls uses the contractarian methodology in a way that is explicitly metaphysical, whereas he is presently inclined to regard metaphysics (and thus the contractarian methodology as I understand it in this essay) as, at the very least, *unnecessary* for the justification of his conception of justice. I am concerned with discussing (only) those contractarian theories that take the idea of a (hypothetical) contract to confer ethical warrant on a moral or political theory.

Mackie – either reject moral objectivism completely or else accept that label for their theories on condition that it be understood in an importantly non-Kantian way. Clearly, there is more than one kind of contractarian moral theory if those who use contract talk in their arguments produce moral theories with such importantly different metaethical foundations.

This essay is interested in critically reflecting on the nature of moral contractarianism. I argue that there are two importantly different kinds of moral contractarian theory in contemporary philosophy, each of which is linked to a different historical figure in the contract tradition, one to Hobbes and the other to Kant. I focus mostly on the Hobbesian variant, recently revived in an exciting form by David Gauthier. I engage in this reflection as one who is deeply attracted to the methodology; I seek to clarify what it is that many of us have found so appealing about the idea that talk of contracts can be a route to understanding the nature of our moral and political obligations.

I

Although Hobbes's masterpiece *Leviathan* is primarily concerned with presenting a contract argument for the institution of a certain kind of state (one with an absolute sovereign), if one looks closely, one also sees a sketch of a certain kind of contractarian approach to morality, which has profoundly influenced contemporary moral theorists such as Gauthier.

Hobbes's approach to morality does not assume there are natural moral laws or natural rights that we discern through the use of our reason or intuition. It is not an approach that assumes there is a naturally good object in the world (such as Aristotle's *Summum Bonum*) that moral action serves and that people ought to pursue. It is not an approach that explains moral action as "natural," for example, as action generated by powerful other-regarding sentiments; Hobbes did not believe that such sentiments were very important or powerful in human life. And it is not an approach that justifies morality as a set of laws commanded by God – although Hobbes believed that his moral imperatives were *also* justified as commands of God.[6] Using his contractarian method, he seeks to define the nature and authority of moral imperatives by reference to the desires and reasoning abilities of human beings, so that regardless

[6] Hobbes believed that moral imperatives were commanded by God, but this justification of them is different from the contractarian justification that he also uses to defend them. The contractarian method seeks to define the nature and authority of moral imperatives by reference to the desires and reasoning abilities of human beings, so that regardless of their religious commitments, all people will see that they have reason to act morally.

of their religious commitments, all people will see that they have reason to act morally. So without repudiating the divine origin of the laws, Hobbes invokes contract language in order to develop an entirely *human* justification of morality.

Without going into a lengthy exegesis of Hobbesian texts, which is bound to be controversial among Hobbes scholars, let me simply state here the features of what I take to be the Hobbesian moral theory. Whether or not every detail of the approach was explicitly embraced by Hobbes himself, its overall structure is recognizably Hobbesian:

1. What is valuable is what a person desires, not what he ought to desire (for no such prescriptively powerful object exists); and rational action is action that achieves or maximizes the satisfaction of desire (where it is a fact that the desire for self-preservation is our primary desire, and that human beings are, by and large, mutually unconcerned).
2. Moral action is rational for a person to perform if and only if such action advances his interests.
3. Morality is, in part, a body of causal knowledge about what human actions lead to peace, an end which it is common knowledge people desire and which they can all share, so that such actions are rational for them and "mutually agreeable." (This precept rests on the Hobbesian belief that people are not self-sufficient, and that they are roughly equal in strength and mental ability.)
4. Peace-producing action is only individually rational to perform (hence only moral action) when there is a convention in the community that people perform such action (so that I know that if I behave cooperatively, then others will do so too, and vice versa). These conventions comprise the institution of morality in our society. The rationality of performance is, however, subject to two provisos:

   Proviso 1: In order to be moral, an action must be not only peace producing and performed in the knowledge that others are willing to do so, but also an action that involves no net loss for the agent.[7]

   Proviso 2: Human beings are not, as a group, rational enough to be able to institute moral conventions, and hence must create a sovereign who can use his power to generate them.

[7] In his book *Hobbesian Moral and Political Theory* (Princeton: Princeton University Press, 1986), Gregory Kavka denies that Hobbes held this proviso, arguing that Hobbes was a "rule egoist."

5. Defining justice or equitable treatment in situations of conflict is done by considering what principles of justice the people involved "could agree to" or "what they would be unreasonable to reject," where the reasonableness of rejection is determined by a calculation comparing the benefits and costs of accepting an arbitrator's resolution with the benefits and costs of resorting to violence to resolve the conflict. An impartial judge, therefore, arbitrates according to the principle "to each according to his threat advantage in war."

Note that contemporary contractarians have enthusiastically taken on the role of Hobbes's arbitrator, trying to determine the sort of division of goods that could be accorded people in a peaceful way, such that they would perceive the division as fair. Nonetheless, we shall see that most of them, including Gauthier, reject the idea that the "threat advantage" of the parties in this situation is defined by their ability to fight.

Let us reflect, for a moment, on the interesting features and strengths of a moral theory with this structure. Consider, first of all, that the Hobbesian approach relies on a very strong conception of individuality. According to Hobbes, cooperative social interaction is presented neither as inevitable nor as something that people value for its own sake, but rather as something that asocially defined individuals find instrumentally valuable given their primary (nonsocially defined) desires. To think that cooperative behavior needs to be encouraged and justified, so that we must be *persuaded* to behave socially toward one another, is to believe that, even if society has some affect on us, it does not determine our fundamental or "intrinsic" nature as human beings, which is a nature that "dissociates us, and renders us apt to invade and destroy one another" (*Leviathan* 13, 10, 62).

Moreover, notice that there are two quite different ways in which this moral contractarian theory uses the notion of agreement. Features 2 and 3 capture the idea that the behavior enjoined by Hobbes's laws of nature is "agreeable," that is, that such action helps to secure the most-desired objects and/or states of affairs for each individual. Feature 5 captures the idea for which moral contractarians are famous; namely, that certain features of morality (e.g., fair resolution of conflict) can be understood as the *object* of agreement. However, there is a connection, in Hobbes's theory, between the latter way of using agreement and the former. To resolve conflicts via the use of arbitrators and agreement procedures is to resolve them peacefully and with much less cost to the parties than more violent resolution procedures. Hobbes commends the use of arbitrators as individually rational for disputants, and warns the arbitrators that their usefulness to the disputants depends on the extent to which their peaceful resolution is more acceptable than going to war to resolve

the dilemma. It is therefore conducive to self-preservation to use a cooperative agreement procedure to resolve conflict, so that defining moral behavior through agreement is itself, for Hobbes, a mutually agreeable – that is, mutually self-preserving – behavior.

But perhaps most important of all, we should appreciate that all five features of Hobbes's moral view fit into a moral theory that is committed to the idea that morality is a *human-made institution*, which is justified only to the extent that it effectively furthers human interests. That is, Hobbes seeks to explain the *existence* of morality in society by appealing to the convention-creating activities of human beings, while arguing that the *justification* of morality in any human society depends upon how well its moral conventions serve individuals' desires.

In fact, there is a connection between Hobbes's contractarian approach to the state and this approach to morality. His decision to justify absolute sovereignty by reference to what people "could agree to" in a prepolitical society is an attempt to explain and legitimate the state's authority by appealing neither to God nor to any natural features of human beings that might be thought to explain the subordination of some to others, but solely to the needs and desires of the people who will be subjects of political realms. In the same manner, he insists that existing moral rules have power over us because they are social conventions for behavior (where Hobbes would also argue that these conventions only exist because of the power of the sovereign).

But Hobbes does not assume that existing conventions are, in and of themselves, justified. By considering "what we *could* agree to" if we had the chance to reappraise and redo the cooperative conventions in our society, we are able to determine the extent to which our present conventions are "mutually agreeable" and so *rational* for us to accept and act on. So Hobbes's moral theory invokes both actual agreements (i.e., conventions) and hypothetical agreements (which involve considering what conventions would be "mutually agreeable") at different points in his theory; the former are what he believes our moral life consists of; the latter are what he believes our moral life *should* consist of – that is, what our actual moral life should model. The contractarian methodology is useful in defining and justifying morality for one who believes that morality is man-made because considering what moral laws "people could agree to" (as well as what laws they have agreed to) is a way of confirming *that* morality is man-made, and a way of appraising how well the present institution serves the powerful self-regarding interests that virtually all of us have.

Note that this way of cashing out the language of hypothetical agreement makes the agreement-talk only a kind of metaphor, and not a device that reveals, in and of itself, the nature of morality or justice. What rational agents could all agree to is the securing of an object and/

or state of affairs, the benefits of which they could all share and for which there is a rational argument using premises that all rational agents would take as a basis for deliberation. Hence, to determine what these agents "could all agree to," one must perform a deduction of practical reason, something that Hobbes believes he has done in Chapters 14 and 15 of *Leviathan*.

Hence, the notion of contract or agreement does not do justificational work *by itself* in the Hobbesian moral theory. What we "could agree to" has moral force for Hobbes not because make-believe promises in hypothetical worlds have any binding force, but because this sort of agreement is a device that *reveals* the way in which the agreed-upon outcome is rational for all of us. The justificational force of this kind of contract theory is therefore carried within, but derived from sources other than, the contract or agreement in the theory.

II

There was enormous interest in this Hobbesian understanding of morality in the seventeenth century by both detractors and supporters alike.[8] The theorist who did most to advance this Hobbesian moral project before the twentieth century in what he took to be the right, and more plausible, direction was David Hume, a philosopher who tends to be incorrectly classed with the Benthamite utilitarians who followed him.[9] Hume was far more willing than Hobbes to credit people with substantial other-regarding desires that he considered to be the (natural) source of many moral virtues, but he also insisted that our self-regarding desires could cause us to invent "artificial" virtues such as being just, respecting others' property rights, keeping one's promises, and being chaste. Hume is so clear in presenting the creation of these virtues as a conventional solution to a coordination problem that David Lewis uses Humean remarks and examples to illustrate what coordination problems are and how conventions work to resolve them.[10]

In the latter half of the twentieth century, we find renewed enthusiasm for this approach and a sustained interest in developing it further. And I suspect that the source of the enthusiasm comes from contemporary

[8] Clearly, Spinoza was influenced by it, but lesser-known theorists found it just as intriguing and some actually sought to "derive" morality from self-interest in Hobbesian fashion even while vociferously denouncing Hobbes's political conclusions and some of his moral laws – for example, Richard Cumberland and Locke's friend James Tyrell. Some unpublished fragments even suggest that Locke himself was intrigued by this project (although the *Second Treatise* assumes that morality is composed of God-made and God-justified natural laws).

[9] See David Gauthier, "David Hume: Contractarian," *Philosophical Review* 88(1) (1979): 3–38, for a discussion of why and how Hume should be understood as a contractarian.

[10] David Lewis, *Convention* (Cambridge, MA: Harvard University Press, 1969).

philosophers' attraction to the most important and fundamental feature of this approach, the presumption that morality is a human creation. For example, James Buchanan insists that

> Precepts for living together are not going to be handed down from on high. Men must use their own intelligence in imposing order on chaos, intelligence not in scientific problem-solving but in the more difficult sense of finding and maintaining agreement among themselves.[11]

And J. L. Mackie is especially disparaging of the idea that morality is something objective and "out there" that we find rather than create and that exerts inexplicable prescriptive power over us. Instead, Mackie wants us to understand and *use* the fact that morality is something we generate in order to serve our interests, an idea suggested by the very title of his book, *Ethics: Inventing Right and Wrong*.[12] Declaring in Chapter 5 that "Morality is not to be discovered but to be made: we have to decide what moral views to adopt, what moral stands to take,"[13] Mackie goes on to discuss rather briefly the sorts of game-theoretic situations that can make agreement on moral behavior advisable,[14] and he concludes by insisting that insofar as we make morality because of our interests, we might have to remake it, at least in part, when those interests change. The same rebellious overtones of the social-contract argument welcomed by Locke and Hume in a political context are welcomed by Mackie in a moral context. Given that the contract argument presents the state as a human creation designed to serve human interests, it justifies the people's replacing a ruler who fails to serve those interests. Similarly, given that the argument presents morality as a human invention designed to serve human interests, it justifies replacing a moral virtue that has outlived its usefulness.

The contemporary theory that most completely realizes the Hobbesian approach and that develops it in important ways is presented by David Gauthier in *Morals by Agreement*,[15] where he attempts to "validate the conception of morality as a set of rational, impartial constraints on the pursuit of individual interest."[16] Every one of the features of Hobbes's moral theory is embraced in some fashion by Gauthier. On his view, moral behavior is rational and mutually advantageous behavior (features 1 and 2) that will lead to a cooperative state of affairs that is desired by everyone (feature 3), assuming, of course, that they are equal in rationality and technology, when (and only when) people become disposed

[11] James Buchanan, *The Limits of Liberty: Between Anarchy and Leviathan* (Chicago: The University of Chicago Press, 1975), p. xx.
[12] John Mackie, *Ethics: Inventing Right and Wrong* (New York: Penguin, 1977).
[13] Ibid., p. 106.
[14] Ibid., p. 123ff.
[15] David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986).
[16] Ibid., p. 6.

to engage in such behavior on a widespread basis (i.e., when a convention to behave cooperatively exists – feature 4). Gauthier also argues that resolution of conflict by such individuals should proceed via principles arrived at by considering the outcome of a hypothetical bargain among equals (feature 5). The people in this theory are quite clearly determinate individuals, who are defined prior to the morality that their contractual agreement is supposed to justify. While Gauthier does not explicitly say, as Mackie does, that the constraints traditionally endorsed as "moral" in human societies are human inventions, that idea, as well as the idea that these constraints can be "reinvented" to better serve human purposes, appears to be the assumption behind his philosophical project, which aims to show what conventions people *would* agree to if they were the sort of perfectly rational people we are all striving to become.[17]

However, what makes Gauthier's moral contractarianism so interesting is the way in which it develops certain features of Hobbes's moral theory to produce not only a more sophisticated moral theory than Hobbes's own, but also one that is more palatable to twentieth-century moral theorists. Consider again feature 4 of Hobbes's theory: that it would only be rational to act cooperatively if others are disposed to do so. In general, Hobbes seems to be right that cooperative situations have a game-theoretic structure such that people are rational to act cooperatively together, but irrational to act cooperatively alone. Yet sometimes cooperation is surely going to have a Prisoner's Dilemma structure, so that even when others are disposed to cooperate, the individual agent is still rational *not* to cooperate. This suggests that the correct moral attitude is one that says, in essence: "I will cooperate with others, when they are willing to do so, except in situations where, by not cooperating, I can gain benefits from them with impunity," but this attitude is hardly what one would call "moral." Hume explicitly worries about this problem when he discusses the "sensible knave" who has exactly the attitude I have just described:

> And though it is allowed that, without a regard to property, no society could subsist; yet according to the imperfect way in which human affairs are conducted, a sensible knave, in particular incidents, may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any considerable breach in the social union or confederacy. That *honesty is the best*

[17] For example, see ibid., p. 168: "We do not of course suppose that our actual moral principles derive historically from a bargain, but in so far as the contraints 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"; or page 231ff., where he says that although unequal possession of technology means that the equality assumption does **not** hold in our world (so that the conventions he argues for may not be applicable), we are "moving towards" the equal possession of this technology, and so, of equal rationality.

policy, may be a good general rule, but is liable to many exceptions; and he, it may perhaps be thought, conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions.[18]

The knave is essentially saying that he will cooperate if and only if it is utility maximizing for him to do so, and thus will be prepared not to do so in situations, such as the Prisoner's Dilemma, despite the existence of a moral convention to perform the cooperative act in that sort of situation. And what does Hume say to this sensible knave? Essentially nothing. Given the difficulties that Hobbes himself had providing an answer to the same knavish question, we see that it is difficult for anyone who embraces the Hobbesian approach to morality to persuade someone who has no natural sentiments against exploitation of his fellow man not to exploit them when he can do so with impunity. Yet such a person is very far from being moral.

Gauthier attempts, however, to answer the knave, inspired by a line of argumentation that he believes Hobbes suggests (but does not develop adequately) in an attempt to answer the "foole" – who offers roughly the same challenge as Hume's knave.[19] It is rational, says Gauthier, for people to become "disposed" to cooperate in such situations (assuming, however, that a sufficient number of others will become similarly disposed). By doing so, they become "constrained maximizers" rather than knavish "straightforward maximizers," where the former are people who pursue their advantage but who do so respecting a constraint against exploitative noncooperation in Prisoner's Dilemmas, where they have good reason to believe that their partners are inclined to cooperate.[20] Such people are willing to forego benefit in Prisoner's Dilemmas; hence, they are not straightforwardly maximizing utility. Yet they have chosen to be disposed to act in this way because they have determined that they can amass more utility by having this disposition than by not having it. A constrained maximizer refrains from taking advantage of any person who is also disposed to constrain his maximizing behavior because "he is not the sort of person that is disposed to do that sort of thing." That is the "moral" attitude that the sensible knave lacks. But the constrained maximizer has that "moral" attitude because of a prior determination that it is individually utility maximizing to have it. So, true to Hobbesian principles, Gauthier is arguing that moral behavior

[18] David Hume, *Enquiry Concerning the Principles of Morals*, edited by L. A. Selby-Bigge, revised by P. H. Nidditch (Oxford: Clarendon Press, 1975), sec. ix, pt. ii, pp. 281–3.

[19] See *Leviathan*, chap. 15. Hobbes's answer to the fool is much discussed in the secondary literature. I give a review of recent interpretations, and one of my own, in *Hobbes and the Social Contract Tradition*, chapters 2 and 3. Gauthier discusses the passage in various parts of chapter 6 of *Morals by Agreement*.

[20] Gauthier, *Morals by Agreement*, p. 170.

is utility maximizing and, in the long run, behavior that involves no net cost.

Contemporary Hobbesians and Humeans would certainly *want* to embrace Gauthier's argument if they could. It offers them a way to explain how collectively rational cooperative action that involves forgoing exploitative opportunities, but which is not dangerous, is also *individually rational* for the agent. But I am not so sure that they can embrace it. First, the idea that one could "will" to be disposed to act as Gauthier describes is dubious if one accepts Hobbesian psychology, and perhaps just as dubious on more plausible contemporary psychological theories. Second, it remains to be seen whether or not Gauthier's argument that it is rational to become disposed to act as a constrained maximizer actually succeeds. If Peter Danielson is right (see his essay, chap. 16 in this volume), it is rational to adopt the more "knavish" cooperative attitude called "reciprocal cooperation," which differs from Gauthier's constrained maximization in that it directs us to exploit (rather than cooperate with) unconditional cooperators. Finally, it might be even more rational only to *pretend* to be disposed to cooperate in either Gauthier's or Danielson's sense, ready to exploit others whenever one can do so with impunity.

The jury is, therefore, still out on the question of whether constrained maximization is rational for individuals to adopt. But other of Gauthier's modifications of Hobbes's project face what seem to be even more serious difficulties. For example, Gauthier argues that Hobbes was wrong to think that we could not establish moral conventions voluntarily, and that we need a sovereign to make their creation possible (although he admits that a limited political power would be needed to handle those among us who are not rational). Not only are most people able to constrain their maximizing tendencies for long-term gain on his view, but they are also able to recognize and act from a principle of acquisition that will provide a rational starting point for further agreement on the terms of cooperation. This principle is what Gauthier calls the "Lockean Proviso" – which directs that one is to acquire goods in a way that leaves no one worse off; and the principle defining fair terms of cooperation that rationally proceeds from a bargain based on this proviso is what Gauthier calls the principle of "minimax relative concession" (hereafter the MRC principle), which essentially directs that the parties are to accept that outcome that is the result of their making equal concessions to one another in the bargaining process.

It may appear that Hobbes has no equivalent of the proviso or the MRC principle since, in his view, there is no way that people could develop a peaceful method of acquiring or dividing goods outside of civil society. But this is not quite so; as we saw, he does consider the kind of principle that an arbitrator (were such a thing possible in

the state of nature) would be rational to use in resolving disputes about the acquisition or the division of goods: "to each according to his threat advantage in war." Clearly, there is a big difference between this principle and Gauthier's cooperative rules! Is either theorist's argument for his approach effective?

James Buchanan comes down on the side of Hobbes. Imagine, says Buchanan, a state of nature in which people are competing for some scarce good x:

> Each would find it advantageous to invest effort, a "bad," in order to secure the good x. Physical strength, cajolery, stealth – all these and other personal qualities might determine the relative abilities of the individuals to secure and protect for themselves quantities of x . . . as a result of the actual or potential conflict over the relative proportions of x to be finally consumed, some "natural distribution" will come to be established.[21]

It is this "natural distribution" that then becomes the baseline for any further contractual agreements. And it is that distribution that then "defines the individual" for purposes of future bargaining.

This future bargaining should occur, according to Buchanan, because everyone has a motive for resolving disputes and allocating goods peacefully given the substantial costs of predation and defense. Successful resolution of conflict through peaceful means would free up the resources used in warfare, and any agreement reached regarding the distribution of these resources, or any portions of the good x not appropriated, will proceed from the natural distribution. What Buchanan does not notice is that the natural distribution also generates the principle to be used in the peaceful resolution of these sorts of competitive conflicts: it is the principle "to each according to what he would have received in war."[22] Consider the following passage from *Leviathan*:

> if *a man be trusted to judge between man and man*, it is a precept of the law of nature, *that he deale Equally between them*. For without that, the controversies of men cannot be determined but by War. He therefore that is partial in judgement, doth what in him lies, to deter men from the use of Judges, and Arbitrators; and consequently (against the fundamental Law of Nature) is the cause of War. (*Leviathan* 15, 23–4, 77)

Hobbes is saying here that an arbitrator in a dispute must beware not to be "partial" in his resolution of the conflict or else the parties will ignore his resolution and go to war to resolve their dispute. But the knowledge that warfare may be deemed rational by the parties if the outcome is not to their liking will affect how the arbitrator resolves

[21] Buchanan, *The Limits of Liberty*, p. 23.

[22] I am indebted to a discussion of Hobbes on justice by Brian Barry at the 1986 Pacific Division Meeting of the American Philosophical Association for the kind of strategy I am using to interpret the following passage.

the conflict. He must try, as far as possible, to mimic the distribution of the goods or the resolution of the conflict that the parties believe warfare between them will likely effect (assuming that each would stop short of attempting to kill the other). To do otherwise would be to risk one party deciding, "I won't accept this resolution: I can get more if I go to war." Of course, there are costs to going to war that are not involved in accepting an arbitrator's resolution of the conflict, so that even if the arbitrator got the resolution wrong, he might be close enough to the division each thinks warfare would effect such that no party would feel it was worth the cost of warfare to try to get more. On the other hand, one or both of them may be vainglorious and believe (falsely) that he can win a fight over the other and wrest away everything that he wants, in which case there is no way the arbitator can resolve their dispute that both will find acceptable. But when both are at least fairly realistic in assessing their powers, the arbitrators can peacefully decide conflicts between them using the maxim "To each according to his threat advantage in a conflict between them."

Gauthier argues that the initial bargaining position is misidentified with the noncooperative outcome, and although his argument is directed at Buchanan, it would clearly apply to Hobbes as well. Why, Gauthier asks, should people behave in a way that maintains the effects of predation after it has been banned?

> 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.[23]

Gauthier is, I believe, trying to make the following point. If people have decided to enter a world in which their interactions are cooperative rather than coercive, then coercive power and the goods that this power has amassed no longer define the parties' bargaining positions; instead, it is their power as cooperators that determines their clout in the bargain, as the MRC principle is meant to represent. If Buchanan and Hobbes reply that past coercers can threaten a return to predation and warfare unless they get them, then Gauthier will counter that such a return is extremely expensive for them, so expensive that it would be a threat they would never feel they could carry out. Not only would they lose the resources that had been freed up by the ban on predation, but they would also give up any productive returns that those freed-up resources may have been able to yield in cooperative investments with others. Gauthier argues that distribution according to predative power should be abandoned, and that initial distribution rationally proceeds according

[23] Gauthier, *Morals by Agreement*, p. 196.

to the Lockean Proviso, while the results of further cooperation should be distributed by the market or, when the market fails, according to the MRC principle.[24]

But Buchanan and Hobbes can defend their claim that predative power should still be understood as the foundation of the parties' bargaining on distribution of a cooperative surplus. Imagine a world in which predation has gone on for some time. The predators would certainly prefer to the MRC principle the Hobbesian "warfare threat advantage" principle, which would give every party at least what she would have gotten in the state of war, plus some of the resources that previously went into predation and defense. The predators would point out that no one would lose, and everyone would gain, from this deal, although the weak would not gain as much from this principle as they would from MRC.[25] But why should the weak, who may have considerable cooperative potential, go along with this deal? Doesn't such potential generate a new threat advantage, so that the result of the agreement will be (loosely) "To each according to his production in the cooperative endeavour"? I want to propose that the strong may have a strategy for ensuring that it does not by invoking the very notion of commitment that Gauthier himself thought so powerful in his answer to the knave. The strong would be rational to turn the situation into a two-move game and use what game theorists call a "precommitment strategy," which is essentially just the same as Gauthier's technique of "constraining oneself for gain."[26] On the first move they would perform two actions. They would:

[24] I have argued that bargaining clout derived from cooperative power in Gauthier's contractual situation does not result in selection of the MRC principle, but I want to put that argument aside here. See my "Can We Agree On Morals?" *Canadian Journal of Philosophy* 18 (1988): 331–56, reprinted in part in chap. 10 in this volume.

[25] The rational principle of distribution may be slightly more complicated than this. *If* it were possible to determine which resources would have been put to use by individuals for predation and defense, then these resources would not have to be distributed according to the threat-advantage principle on the grounds that *everyone* would benefit from any principle that gave them a portion of these goods (since in the state of nature, they would get none of them). Moreover, the powerful could not threaten to use warfare to force the others to accept the threat-advantage principle, since this would mean losing the very benefits that were being contested. This appears to be the core of truth in Gauthier's argument. However, a large family of principles would be prima facie acceptable to all of the parties in this situation and the threat-advantage principle would be among them and would certainly have the advantage of being *salient*. Nonetheless, if the isolation of goods that otherwise would have gone to predation is difficult or impossible, than the warfare threat-advantage principle is a rational principle of division for all contested goods. I am indebted to conversations with Stephen Munzer, which helped me to clarify this point.

[26] Hence, I am essentially arguing that if Gauthier's kind of constraining device is psychologically possible for human beings, then this may have bad consequences for other elements of his contractarian theory. I am indebted to Christopher Morris for this way of characterizing my argument.

(a) make a threat to reassemble the means of war for as long as it took to persuade the noncompliers to go along with the threat-advantage principle;

and then they would

(b) dispose themselves to keep their threats, no matter how expensive it is to do so.

The second move would be made by the weak. What is their rational response to the first move of the strong? Clearly, they would find it utility maximizing to accept the threat-advantage principle rather than to hold out for a principle more favorable to them. Hence, by transforming the situation into a two-move game and using the first move to make a threat that they would then commit themselves to keep, the strong would be able to insist on ensuring that the structure of future mutual cooperation respects their past predative power.

In response, Gauthier could try to contend that this kind of two-move strategy would be unavailable to the people in his bargaining situation. For example, he might argue that insofar as the weak are perfectly rational, they would know that this strategy would be rational for the strong, and would do their best to block it.[27] But precisely because they *are* weak, blocking this strategy might be difficult. Indeed, it is difficult for Gauthier to *prove* that the weak could block it. His bargaining situation is so sparsely described and highly idealized that we can find nothing in the structure of that situation to rule out this kind of precommitment strategy by the strong, so that it seems possible for both the starting point and the results of a Gauthierian initial contract to be alarmingly Hobbesian.

These remarks make me appear strangely unappreciative of Gauthier's attempt to mount a plausible neo-Hobbesian moral theory. It seems that I am commending to contemporary contractarians the meanest and most unappealing aspects of Hobbes's approach to justice and property. But those mean and unappealing aspects are quite clearly and strongly linked with Hobbes's requirement that moral action involve no net loss to the agent. There are no free giveaways or free rides on Hobbes's theory; you get what it is in your interest to get and what it is in others' interest to let you have. The results of this kind of thinking are not, I think, very attractive. Contemporary Hobbesians like Gauthier try to accept the self-interested underpinnings of the theory but dress up or deny the conclusions that Hobbes claims they force one to draw. I have attempted to suggest in these remarks that Hobbes is right to insist on them. I

[27] This response to my argument was made on Gauthier's behalf by Geoffrey Sayre-McCord, who commented on this paper at the conference for which this paper was solicited.

suspect that if Gauthier or other theorists sympathetic to the structure of Hobbesian theory long for "nicer" principles of morality and justice than those that Hobbes develops, they need to find a non-Hobbesian foundation for them. And as I now discuss, there are signs that Gauthier himself suspects this is so.

III

Consider what many have found a particularly ugly side to Hobbesian morality: its radical individualism. Recall that the people in Hobbes's or Gauthier's contracting world are fully developed, asocially defined individuals. But when *Leviathan* was originally published, some readers were shocked by the idea that the nature of our ties to others was interest-based. Aristotelian critics contended that Hobbes's theory goes too far in trying to represent us as radically separate from others. Their worries are also the worries of many twentieth-century critics. Do not our ties to our mothers and fathers, our children and our friends define, at least in part, who we are? Isn't it true that our distinctive tastes, projects, interests, characteristics, and skills are defined by and created within a social context? So how can a moral theory that does not take this into account be an accurate representation of our moral life? It would seem that we *must* bring into our moral theory noninstrumental ties with others that are not based on our affections because it is through such ties that we *become* individuals. This is the kind of criticism that certain feminist writers have made of contractarian theory (e.g., Carole Pateman), and it is part of the reason for its rejection by Marxist thinkers (e.g., Macpherson), who regard the individualism in contractarian thinking as a reflection of the theory's bourgeois, capitalist origins.[28]

Hobbes would either not understand or else resist the claims of our social definition. But Gauthier, a member of our place and time, accepts them, and this has strange consequences for his moral theory. Gauthier is moved by the criticism that it is unfair to use allocation procedures, such as the market, to distribute goods in circumstances where the society permits – even encourages – one class of people to prevent development in another class of people of those talents that allow one to do well in a system using that allocation procedure. Thus, he suggests that we see his contract on the fair terms of cooperation not as an agreement among determinate, already defined, individuals, but as an agreement at a hypothetical "Archimedean Point" among "proto-people" – people who have a certain genetic endowment and who are concerned to select principles that will structure their society such that they will develop well:

[28] Gauthier himself has been moved by these kinds of worries, inspired, he says, by Hegel. See his "Social Contract as Ideology," *Philosophy and Public Affairs* 6 (1977): 130–64.

> The principles chosen from the Archimedean point must therefore provide that each person's expected share of the fruits of social interaction be related, not just to what he actually contributes, since his actual contribution may reflect the contingent permissions and prohibitions found in any social structure, but *to the contributions he would make* in that social structure most favorable to the actualization of his capacities and character traits, and to the fulfillment of his preferences, provided that this structure is a feasible alternative meeting the other requirements of the Archimedean choice. (My emphasis)[29]

No longer does Gauthier's contract talk presume fully determinate individuals, and no longer is the object of any contract a principle for the resolution of conflict among individuals. Now the contract methodology is used to choose principles that are "for" the structuring of the social system that plays a profound role in structuring individuals. Like Rawls, Gauthier is declaring that the first order of moral business is the definition of social justice.

This is not a benign addition to Gauthier's Hobbesian moral theory: it is an addition that essentially destroys its character as a Hobbesian theory. Of course, it undermines the individualism of the original Hobbesian theory; many will think that this is no great loss. But it was that individualism that much of the rest of the theory presupposed. Consider, for example, that a Hobbesian theory answers the "Why be moral?" question with the response, "Because it is in your interest to be so." But that answer no longer makes sense in a contract theory designed to pursue the nature of social justice using protopeople. Suppose the results of that theory call for a more egalitarian distribution of resources and opportunities open to talents that society will attempt to develop in all its members. If I am a white male in a society that accords white males privileged opportunities to develop talents that will allow them to earn well, then why is it rational for me to pursue a restructuring of social institutions in which this is no longer true?

Indeed, given that their development has already taken place, *why is it even rational for adult minority members or females to support this restructuring?* All of these people are already "made." Restructuring the social world such that it does a fairer job of creating a future generation of individuals is a costly and other-regarding enterprise. Why should these determinate individuals be rational to undertake it, given its cost, unless they just happened to be affected by sympathy for other members of their race or caste or sex, and so enjoyed the struggle? But the nontuistic perspective Gauthier encourages his bargainers to take encourages them to discount any benefits to others from their actions. So assuming the Hobbesian/Gauthierian theory of rationality, what it would be rational

[29] Gauthier, *Morals by Agreement*, p. 264.

for "proto-me" to agree to in some extrasocietal bargain seems to have little bearing on what it is rational for "determinate-me" to accept now.[30]

It is because the self-interest of *determinate* individuals does not seem sufficient to explain the commitment to the results of a bargain among *protopeople* that one wonders whether Gauthier's eventual interest in defining fair principles for the development of individual talents in a social system betrays a commitment to the intrinsic value of the individuals themselves. And it is the idea that individuals have intrinsic value that is missing from the Hobbesian approach. It has not been sufficiently appreciated, I believe, that by answering the "Why be moral?" question by invoking self-interest in the way that Hobbes does, one makes not only cooperative action, but the human beings with whom one will cooperate merely of *instrumental value;* and this is an implicit feature of Hobbes's moral theory that is of central importance. Now Hobbes is unembarrassed by the fact that in his view, "The *Value,* or WORTH of a man, is as of all other things, his Price; that is to say, so much as would be given for the use of his Power: and therefore is not absolute; but a thing dependent on the need and judgement of another" (*Leviathan* 10, 16, 42). But this way of viewing people is not something that we, or even Gauthier, can take with equanimity. In the final two chapters of his book, Gauthier openly worries about the fact that the reason why we value moral imperatives on this Hobbesian view is that they are instrumentally valuable to us in our pursuit of what we value. But note *why* they are instrumentally valuable: in virtue of our physical and intellectual weaknesses that make it impossible for us to be self-sufficient, we need the cooperation of others to prosper. If there were some way that we could remedy our weaknesses and become self-sufficient, for example, by becoming a superman or superwoman, or by using a Ring of Gyges to make ourselves invisible and so steal from the stores of others with impunity, then it seems we would no longer value or respect moral constraints because they would no longer be useful to us – unless we happened to like the idea. But in this case sentiment, rather than reason, would motivate kind treatment. And without such sentiment, people would simply be "prey" for us.

Even in a world in which we are not self-sufficient, the Hobbesian moral theory gives us no reason to respect those with whom we have no need of cooperating, or those whom we are strong enough to dominate, such as old people, or the handicapped, or retarded children

[30] This point can be made in a slightly different way. If I am a person whose development has in some way been stunted by discriminatory social practices and institutions, I can be the sort of person who will honestly believe that changing the social world such that people of my type receive more and better opportunities is a bad thing for all of us. Moreover, I may be right that given who *I* am, increased opportunities would be a bad thing for *me.* (Consider women who vigorously fight the equal-rights amendment.)

whom we do not want to rear, or people from other societies with whom we have no interest in trading. And I would argue that this shows that Hobbesian moral contractarianism fails in a very serious way to capture the nature of morality. *Regardless* of whether or not one can engage in beneficial cooperative interactions with another, our moral intuitions push us to assent to the idea that one owes that person respectful treatment simply in virtue of the fact that he or she is a *person.* It seems to be a feature of our moral life that we regard a human being, whether or not she is instrumentally valuable, as always intrinsically valuable. Indeed, to the extent that the results of a Hobbesian theory are acceptable, this is because one's concern to cooperate with someone whom one cannot dominate leads one to behave in ways that mimic the respect one ought to show her simply in virtue of her worth as a human being.

Hobbesians themselves are not immune to this moral pressure. Although Gauthier wards it off throughout much of his book, I have argued that he gives way to it when he advocates social justice. I suspect it may also be behind his curious reluctance to admit the legitimacy of using force and coercion in the division of goods despite the fact that the Hobbesian foundations of his theory pressure him to do so. Indeed, signs of respect for persons even break through the surface of the moral theory of that hopeless curmudgeon Hobbes himself! Consider that there is an insistence in *Leviathan* that human beings are *equal;* and it is their equality coupled with their lack of self-sufficiency that make cooperative action instrumentally valuable for them. Yet it is hard to claim that such equality is a *fact;* the world gives us too many counterexamples, and Hobbes himself has difficulty sustaining belief in this "fact." At one point he contends:

> If Nature therefore have made men equal, that equality is to be acknowledged; or if Nature have made men unequal, will not enter into conditions of Peace, but upon Equall terms, such equality must be admitted. And therefore for the ninth law of Nature, I put this, *That every man acknowledge others for his Equal by Nature.* The breach of this Precept is *Pride.* (*Leviathan* 15, 21, 77)

This passage is puzzling. If we really *are* equal, then it makes sense to acknowledge it. And if we are not, why must we pretend to be? Hobbes suggests that warfare will result if we do not. But if I am *genuinely* superior to you, then I should be able to win against you in the battlefield, such that it is not in your interest to fight me. And as Buchanan suspected, my superiority in warfare should mean that I may well be able to succeed in getting you to enter into conditions of peace on unequal terms. Moreover, if I am vastly your superior, I need not enter into *any* cooperative agreement to secure your help for me. I can simply coerce it from you; so the cooperative laws of nature are irrelevant to me.

Hobbes persistently refuses to say such things, and his commitment to either the fact of, or the need for the belief in, equality begins to look suspiciously like faith in a premiss that is required if his moral laws are *always* going to be applicable to us. One suspects that like Gauthier, Hobbes does not want to accept the fact that his theory directs us to find others valuable only when we do not have sufficient strength to dominate them or when we find their help necessary to achieve our own plans. Both philosophers feel the pressure of the idea that respect for persons is something we are compelled to pay no matter what their "price" or power.

IV

To abandon the idea that the only value human beings have is instrumental is to abandon the Hobbesian approach to morality. Gauthier drifted toward this abandonment when he started talking about hypothetical contracts among protopeople, whose value at this point in their nonlives could not be instrumental. This way of talking moved Gauthier in the direction of what I call "Kantian contractarianism."

Consider how Kant talks about the "idea" of the "Original Contract" in the context of defining just political policies:

> Yet this contract, which we call *contractus originarius* or *pactum sociale*, as the coalition of every particular and private will within a people into a common public will for purposes of purely legal legislation, need by no means be presupposed as a fact. . . . It is rather a *mere idea* of reason, albeit one with indubitable practical reality, obligating every lawmaker to frame his laws so that they *might* have come from the united will of an entire people, and to regard any subject who would be a citizen as if he had joined in voting for such a will. For this is the touchstone of the legitimacy of public law. If a law is so framed that all the people *could not possibly* give their consent – as, for example, a law granting the hereditary *privilege* of *master status* to a certain class of *subjects* – the law is unjust. . . ."[31]

For Kant, asking "what people could agree to" when resolving political issues already *assumes* the intrinsic value of each individual, or, as Kant puts it, the idea that they are "ends-in-themselves." And it is a method for ascertaining the fair or just policy. We are to imagine people equally and fairly placed, and we are to give them a veto power over moral proposals put before them. In this way, we represent their value as ends-in-themselves amongst other such ends. So in Kant's view, the contract methodology is a *way of reasoning about how persons with such value ought to be treated.*

[31] Immanuel Kant, "On the Common Saying: 'This May be True in Theory, But It Doesn't Apply In Practice,' " in *Kant's Political Writings*, edited by Hans Reiss (Cambridge: Cambridge University Press, 1970), p. 63.

Kant's use of the social-contract method is picked up by Rawls in *A Theory of Justice*, and used to pursue the nature of social justice that the *fact* of our social nature makes the first item of moral business. Scanlon also makes this kind of use of the contract methodology, seeing it not so much as a tool for social justice, as a general procedure for defining a variety of moral obligations toward others. For these theorists, morality is not so much "invented" as it is "constructed," where the latter term is meant to suggest that specific moral laws and conceptions are not self-interested inventions of human beings, but theorems derived from human reason. For example, Rawls explicitly compares his original position procedure to Kant's categorical imperative procedure,[32] and Scanlon suggests that the contractarian form of argument is a kind of proof procedure for ethics, analogous to proof procedures in mathematics, having its basis in human reason and which we use to construct moral laws in a way that gives them objectivity.[33]

But Hobbesians such as Gauthier would be highly suspicious of the claim that this sort of approach to morality offers a justification of moral behavior. If the method works by *assuming* that human beings have equal, objective, and intrinsic value, they would ask how one could presuppose something as controversial, unsupported, and ill-defined as this. And even if one could make sense of the idea that we are valuable as "ends-in-ourselves" and establish that it was true that we had such value, they would wonder why any of us should *care* about respecting that value in others.

Kantians might reply by maintaining that our nature as ends-in-ourselves is a fact that has enormous prescriptive force. Or they might try to argue that this value should itself have the status of a human invention, a creation of civilized human culture that has instrumental value for its members. This last answer may have some appeal for Hobbesians, but is unlikely to be "objective" enough for those Kantians who want to see our intrinsic value as something more than a useful pretense.

Alternatively, the Kantian might give up the idea that the contract method assumes our value, and argue instead that the method generates the concept of treating human beings as ends. In this view, the method (and its placement of human beings relative to one another) would be taken to be foundational (e.g., as identical to reason in its practical form), and the use of that method to determine how to act in the world with regard to others would, indirectly, be the way one would construct and give content to the concept of treating someone as an end. (So I am treating you as an end if I behave toward you in such and such a way,

[32] Rawls, *Theory of Justice*, sec. 40.

[33] Thomas Scanlon, "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, edited by A. Sen and B. Williams (Cambridge: Cambridge University Press, 1982).

as mandated by the contract method.[34]) But even assuming that the contract method can be understood as conceptually prior to determinations of the value of human beings, the reply seems only to push the problem back a step, for now the Hobbesians will certainly challenge either the existence or the authority (or both) of this supposedly foundational method, and will be poised to attack any account of our motivation to follow it which insists (e.g., as Kant's would) on the motivational force of moral reasoning.[35]

Any of the justificational accounts just presented finds the justificational force of the contract method in sources *other* than the actual contract used in the method, and this was also true (as we discussed) of the Hobbesians' approach to the justificational force of their contract method. However, even if theorists such as Gauthier didn't directly use the contract to provide endorsement of the agreed-upon results, they did use it as a device to reveal the actual justification of those results, that is, their instrumental value. So the contract has a definite use in their theory. But Gauthier would question whether the Kantian's contract talk has any use at all. It is not just that he would question the force of make-believe promises in hypothetical worlds. The problem is more serious. Given that the Kantians traditionally structure the reasoning and information of the bargainers in the hypothetical contract in great detail, it appears they present us with not so much a *bargain* as a theorem in rational-choice theory based on certain supposedly weak and readily acceptable axioms (which generally turn out to by anything but weak and readily acceptable) involving the equality and "fair placement" of the parties relative to one another in a way that represents their equal intrinsic value.

Perhaps Kantians could answer this challenge by resisting any move to make determinate the reasoning of the parties in their arguments. Their hypothetical contract might be best understood not as a theorem in rational choice theory, but as a model that contains ethical intuitions that are too complex and nuanced to be successfully axiomatized, but which are present and alive to us in the right kinds of ways when we attempt to determine an answer to a moral quandary by asking, "What could people agree to – what would they be unreasonable to reject?" Still, such indeterminacy makes the contractarian "proof procedure" a messy, ill-defined business, and so reliant on intuitions as to make it

[34] This reply is suggested by Kant's own characterization of his categorical imperative procedure in a passage of his "Paradox of Method," in the *Critique of Practical Reason*, trans. by Lewis Beck (Indianapolis, IN: Bobbs-Merrill, 1956), p. 65–7.

[35] They would not, presumably, challenge Scanlon's approach to this motivational issue, which involves positing a desire in us to act so as to receive others' endorsement/agreement. But they would probably claim such a desire was not very widespread, or else not very strong, in most of the population.

appear only a more sophisticated variant of traditional ethical intuitionism.

It remains to be seen whether or not Kantian contractarianism can be developed to answer these sorts of problems. I confess to being (perhaps unreasonably) optimistic that it can. However, there is a third kind of criticism that critics of *all* kinds of contractarian theory would want to make of this Kantian variant. Is not the Kantian variant still too individualistic? The Kantian's persistent use of Hobbes's premise that the parties to the contract are "mutually unconcerned" seems to show that even these contractarians are still heavily in the grip of outmoded and distorting individualistic thinking (indicative of the fact that this kind of theory was developed by capitalists and by males).

However, I have concluded after reflecting on certain remarks of Gauthier that Kantian contractarianism is right headed precisely because it preserves this remnant of Hobbesian moral theory, and that its Hobbesian side is what makes it a theory to which feminists (if not Marxists) should be attracted.

Hobbes's central insight about ethics was that it should not be understood to require that we make ourselves a prey for others. It is this insight that both varieties of contractarianism respect. Consider a relationship between two human beings that exists for reasons of either love or duty; let us also suppose that it is a relationship that can be instrumentally valuable to both parties. In order for that relationship to receive our full moral endorsement, we must ask whether either party uses the duty or the love connecting them in a way that affects the other party's ability to realize the instrumental value from that relationship. To be sure, good marriages and good friendships ought not to be centrally concerned with the question of justice, but they must also be, at the very least, relationships in which love or duty are not manipulated by either party in order to use the other party to her detriment. In Gauthier's words, our sociality

> becomes a source of exploitation if it induces persons to acquiesce in institutions and practices that but for their fellow-feeling would be costly to them. Feminist thought has surely made this, perhaps the core form of 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 benefit.[36]

The contractarian should insist that this be true for *any* social relationship, not just that of the polis.

Of course, no contractarian would want to deny that love can lead one to willingly give up one's benefits. It can also lead one to serve others who cannot, for various reasons, reciprocate; for example, infants,

[36] Gauthier, *Morals by Agreement*, p. 11.

or the impoverished, or the aged. Gauthier's remarks suggest not that one should never give gifts out of love or duty without insisting on being paid for them, but rather that one's propensity to give gifts out of love or duty *should not become the lever that another party uses to get one to maintain a relationship to one's cost.*

Perhaps this is most deeply true within the family. Consider a woman whose devotion to her family causes her to serve them despite the fact that they fail to reciprocate; in this case, they are exploiting her love and sense of duty, which cause her to maintain a relationship with them and to serve them. Of course, infants cannot assume any of her burdens; fairness cannot exist between such radically unequal persons. (Note that this relationship is not unfair either; the infant does not use the mother's love in order to exploit her.) But older children can. Indeed, as children grow into the equality they will eventually attain as adults, it is increasingly alarming to see them treating Mom as the maid. Unless they are encouraged to benefit her as they become able to do so, they are being allowed to exploit another human being by taking advantage of her love for them.

Gauthier's remarks suggest that contractarianism of either variety is a moral theory that insists that relationships among equals can only prosper against a backdrop of reciprocity. Our ties to those who are able to reciprocate what we give to them (as opposed to victims of serious diseases, impoverished people, infants) are morally acceptable, healthy, and worthy of praise only insofar as they do not involve, on either side, the infliction of costs or the confiscation of benefits over a significant period of time. What kind of costs and benefits are we talking about here? Not the costs and benefits that come from the affection itself – for among other things, these cannot be distributed and are outside the province of justice, but rather the nonaffective costs and benefits that the relationship itself creates or makes possible.

A person who would inflict such costs on another without compensating him for them or who would take such benefits without benefitting him in turn is failing to respect that person's value and importance as a human being. And the person who would *allow* this exploitation in the name of love or duty is failing to respect his own value and importance. Indeed, if the two of them are supposed to be in a loving relationship (e.g., a marriage), then the exploiter is failing to love the other in virtue of what she does. So love and justice are not opposing responses, for the latter, in the form of reciprocity, is built into the former.

Hence, the Kantian contractarians' insistence on each party's self-concern in his method is not an embarrassment. They need not say (as Hobbes very nearly does) that we are *only* self-interested. Instead, they can be interpreted as saying that in the process of determining whether any loving or duty-based relationship among equals is exploitative, one

must leave aside these connections and ask, from the standpoint of each party, "Is the present distribution of the possible nonaffective costs and benefits of the relationship one which I would be unreasonable to reject?" The Kantian contractarians' insistence that each party to a relationship take such a self-concerned perspective is really the insistence that each of us is right to value ourselves, our interests, and our projects, and right to insist that we not become the "prey" of other parties in the pursuit of their projects. So the contractarian method is valuable in my view not merely because of the way in which it forces us to take into account the well-being of others, but also because it is a method that grants us what Charlotte Brontë in the quotation at the beginning of this essay seems to want, namely, a way to be tenacious advocates of ourselves.[37] What has attracted so many to the contractarian form of argument and what makes it worthy of further pursuit is precisely the fact that by granting to each individual the ability to be her own advocate, this method enables us to conceive of both public and private relationships without exploitative servitude.[38]

[37] Remember that it was because of Rawls' conviction that the utilitarian calculation, which does not incorporate such claims, thereby allowed for the exploitation of some by others that he turned to a contractarian mode of reasoning to get a better understanding of justice. A utilitarian might respond by maintaining that he allows each of us to count equally in the utilitarian calculation. But this way of "counting" still isn't good enough for the contractarian, who would note that each person appears in the utilitarian calculation as a number representing how much he contributes to the total good. This means that it is not **really** the individual so much as the summable units of good that he or she contributes (and, in the final analysis, represents in the calculation) that the utilitarian takes seriously. Each individual is, therefore, rewarded (and valued) by that theory (only) to the extent that he or she responds to any resources by contributing units of good to the total.

[38] I wish to thank Carl Cranor, David Dolinko, Alan Donagan, David Gauthier, Christopher Morris, Stephen Munzer, Geoffrey Sayre-McCord, an anonymous referee for Cambridge University Press, and members of the Saturday Afternoon Discussion Group in Los Angeles during the Spring of 1987 for their comments on, and help during the writing of, this paper. Portions of the paper were read at the 1988 Pacific Division Meeting of the American Philosophical Association.

6. Moral standing and rational-choice contractarianism*

Christopher W. Morris

One of the issues raised by the contemporary debate about abortion concerns the *scope* of our moral principles: to what extent, if any, are human fetuses protected by moral prohibitions of killing and harming? Similarly, there is considerable controversy about the moral status of, for example, humans in irreversible comas, the mentally handicapped, nonhuman animals: to what extent, if any, are they owed moral consideration? Scope questions, or issues about the moral status of particular creatures, are familiar to all conversant with the philosophical literature concerning contemporary moral problems.

In the past, controversies about the scope of moral principles or norms were different. Our ancestors debated the moral status of members of other races and of foreigners or barbarians. These no longer are our concerns, and presumably those of our descendants will differ from ours. The theoretical considerations, however, are the same: namely, what is the moral status of different sorts of creatures or entities? Who, or what, counts? Let us say that to be owed (some) moral consideration is to possess (some) *moral standing*.[1] Then the question is, what has moral standing?

The manner in which this question is usually addressed in the literature is worth noting. It is asked in virtue of what features do entities acquire moral standing. And, typically, the features are held to be certain natural or nonconventional attributes or properties of the entity in question. So, for instance, inquiries are made concerning the rationality, self-consciousness, species membership, or sentience of some creature; and it is usually concluded that some feature (e.g., rationality, sentience, humanity) is either necessary or sufficient (or both) for moral standing. Moral standing is, in this manner, typically ascribed or denied to fetuses, the retarded, nonhuman animals, landscapes, art objects, monuments, and the like.

*For comments on an earlier draft, I am grateful to Peter Danielson, Edward McClennen, Wayne Sumner, and Peter Vallentyne.

[1] The notion of moral standing I borrow and adapt from L. W. Sumner, *Abortion and Moral Theory* (Princeton: Princeton University Press, 1981), p. 26ff.

There are two theoretically significant features of this approach to criteria of moral standing that it is important to highlight, and that contrasts with the approach that I examine in this essay. The first is it is normally assumed that the conditions that contribute to something's possessing moral standing are nonconventional, nonrelational properties or attributes of the creature in question. For instance, rationality, species membership, sentience, and the like are attributes of individuals that are possessed independently of relations to others.[2] And it is the attribute, or cluster of attributes, itself that generates moral standing.

The second significant feature of this approach to criteria of moral standing is that it is implicitly assumed that such standing is universal in a particular way: if something has moral standing, it has it in relation to all moral agents. That is, it is usually assumed that there is one morality that binds all moral agents to accord moral considerations to all who possess moral standing.

Utilitarians thus accord moral standing to all, and only, sentient beings,[3] and members of the natural law and Kantian traditions tend to focus on rationality as the decisive property. By contrast, some moral theories – namely, certain contractarian theories – do not regard the possession of any natural or independent properties as sufficient for moral standing. For such theories, moral standing requires in addition that individuals be related to one another in certain ways. Further, some of these theories even impose certain conditions regarding the *will* of individuals with moral standing; to have moral standing requires that one act toward others in certain ways. Thus, many of these moral theories may deny moral standing to, for example, some rational, sentient humans who do not stand in the appropriate relations to other agents or behave in prescribed ways. Additionally, these theories will tend to be relativist in the sense that they deny that there is a single morality that binds all moral agents to all who possess moral standing. For these theories, moral standing will typically be possessed only in relation to

[2] Or at least of *particular* others. For some philosophers claim that our rationality and self-consciousness are developed, or even constituted, by our relations with others. As will be clear presently, even views of moral standing that employ such "social" accounts of rationality and the like contrast with the view that is examined in this essay.

[3] A utilitarian defense of a sentience criterion is defended by Sumner in *Abortion and Moral Theory*.

proper subsets of moral agents. For these theories, possession of moral standing is a more complicated matter than for the more familiar theories mentioned before.

It is the different and complicated manner in which such contractarian theories accord individuals moral standing that I shall explore in this essay. In particular, I shall focus on the particular account of moral standing implicit in David Gauthier's neo-Hobbesian theory, "morals by agreement." Many find the implications of Gauthier's theory regarding moral standing implausible or at least unsettling. It is important to determine both to what extent these implications do follow from the theory and to determine exactly what assumptions are therefor responsible. Further, it is important to determine to what extent such implications are an unavoidable feature of such moral theories. I shall not, however, do more than determine the account of moral standing implicit in morals by agreement. For many critics, this will suffice, given the counterintuitive nature of some of the implications. This suffices as well for those convinced that contractarianism is the most plausible approach to moral theory, or at least to the theory of justice.

Rational-choice contractarianism

There is a long western tradition, dating back to Glaucon, developed by Hobbes, Hume, and Rousseau, and continued by K. Baier, Rawls, Mackie, Harman, Scanlon, Gauthier, and others[4] that understands morality, or at least justice, to be conventional and, in some sense, the product of agreement. According to this tradition, the norms of morality or justice are conventions that ideally serve to advance the goals of all in certain situations. This tradition is dubbed "contractarian" as it often understands the terms of justice to be the outcome of a hypothetical bargain or "social contract." We may, perhaps less misleadingly, think of contemporary versions of the tradition – especially Gauthier's – as offering a "rational-choice" conception of morality after John Rawls' remark, "The theory of justice is a part, perhaps the most significant part, of the theory of rational choice."[5] We focus on David Gauthier's "morals by agreement."

[4] For the latter, see Kurt Baier, *The Moral Point of View*, abridged edition (New York: Random House, 1958, 1965); John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971); J. L. Mackie, *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin, 1977); Gilbert Harman, *The Nature of Morality* (New York: Oxford University Press, 1977); T. M. Scanlon, "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, edited by A. Sen and B. Williams (Cambridge: Cambridge University Press, 1982), pp. 103–28; and David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986). See also Gregory S. Kavka, *Hobbesian Moral and Political Philosophy* (Princeton: Princeton University Press, 1986).

[5] Rawls, *Theory of Justice*, p. 16. Rawls no longer accepts the view expressed in this quote. See "Justice as Fairness: Political not Metaphysical," *Philosophy and Public Affairs* 14 (1985): 237n.

David Gauthier's theory of morals by agreement is an account of the constraints that agents have reason to accept to bring about the mutually advantageous outcomes not attainable by independent, rational action. These constraints Gauthier identifies with the requirements of morality. Thus, he says, "moral theory is essentially the theory of optimizing constraints on utility-maximization" (*MbA*, p. 78).[6]

Starting with a maximizing account of rationality and a subjective account of value, Gauthier provides an analysis of the problems that rational individuals face in interaction, problems typified by the well-known Prisoners' Dilemma. The core of his theory consists in arguing that, under certain conditions, (1) rational agents faced with such problems will accept constraints on their individual actions, (2) they will rationally comply with these constraints, even in the absence of government or collective enforcement, and (3) these constraints are those of morality, albeit an ideal, as opposed to a common sense, morality. In addition, (4) Gauthier argues for two particular sets of constraints on interaction: (a) a principle that determines the division of the fruits of rational cooperation (the principle of "minimax relative concession") and (b) a "Lockean Proviso" and a consequent set of basic moral rights and duties that define the starting point or baseline for cooperative interaction.

Morals by agreement is what I call, after Rawls' remark, a rational-choice theory of morality or justice.[7] Rational-choice theory is the theory of rationality for individual and social choice, whether in context of independent or of interdependent choice. The conception of rationality invoked and developed is that of utility maximization, where an individual is rational insofar as he or she maximizes the satisfaction of his or her preferences.[8] Rationality and justice are related thus. Rational individuals may often find themselves in situations where cooperative, nonmaximizing behavior is mutually beneficial, given their preferences; these are "the circumstances of justice."[9] Yet in the absence of constraints

[6] References to *Morals by Agreement* (*MbA*) are made parenthetically.

[7] Justice is traditionally understood to be one of many moral virtues. Gauthier's theory focuses on justice, suggesting that for morals by agreement, it is the main, if not the sole, moral virtue. Henceforth, I talk about justice, leaving it open whether there are other contractarian moral virtues.

[8] "Preferences," in this technical sense, are rankings of outcomes. To have an ordering that can be maximized, preferences must be represented by a function that satisfies certain conditions or axioms (e.g., completeness, transitivity). Such preferences, we say, are *coherent*. In addition, for the purposes of moral theory, Gauthier requires that preferences be *considered* (*MbA*, p. 29ff.).

[9] The main conditions are relative but variable scarcity and self-bias (*MbA*, pp. 113–14). The phrase is from Rawls; for his account of these conditions, see *A Theory of Justice*, p. 126ff. See also H. L. A. Hart, *The Concept of Law* (Oxford: Oxford University Press,

on individually rational behavior, such cooperation may not be possible, for such individuals will act in maximizing ways that disadvantage others. Justice imposes the constraints necessary to make mutually beneficial cooperation possible, thereby stabilizing Pareto-efficient outcomes in situations analogous to *n*-person Prisoners' Dilemmas.

Several cooperative outcomes may each be Pareto-efficient. Given that rational individuals will not be indifferent as to which is selected by the norms of justice, we may imagine the specific constraints of justice to be determined by a mutually advantageous agreement to select (principles for the selection of) particular Pareto-efficient outcomes. Cooperation is advantageous to all, but there may be different cooperative arrangements, each distributing the benefits of cooperation differently. We may think of the agreement to select particular forms of cooperation as a type of bargain: each individual presses for the cooperative arrangement most beneficial to him or her, and all agree to some mutually acceptable cooperative arrangement. The theory of justice thus makes use of a particular part of the theory of rational choice, namely, bargaining theory.

How are we to conceive of this bargain, or more importantly – for this will determine what is necessary to possesses moral standing – how are we to conceive of these bargainers? First, they are rational, that is, utility maximizers. Second, they are maximizers of subjective value, for Gauthier takes utility to be a measure of preference and value to be determined solely by (coherent and considered) preference.[10] This has the important consequence that values are relative to individual perspectives, that is, agent-relative.

For the purposes at hand, this may be thought to suffice as a characterization of rational bargainers. For such individuals, whatever their preferences, may find themselves in Prisoners' Dilemmalike situations. All that is needed for such dilemmas is that the preferences of individuals have a certain structure; it is not necessary that they be self-regarding. And agent relativity suffices for preferences to have the requisite structure.[11] It is important, then, to note that the introduction of a utility-

1961), pp. 189–95. The classical accounts are to be found in Hobbes and Hume. For the latter, see esp. *An Enquiry Concerning the Principles of Morals*, sec. III, pt. I, p. 149.

[10] "Value is then not an inherent characteristic of things or states of affairs, not something existing as part of the ontological furniture of the universe in a manner quite independent of persons and their activities. Rather, value is created or determined through preference. Values are products of our affections" (*MbA*, p. 47; see also pp. 24–6).

[11] Gauthier writes:

> For the basic contractarian argument, that it is advantageous for each person to comply with constraints that it would be rational for all to agree to, provided others may be expected to be generally similarly compliant, does not depend in any way on supposing that persons have nontuistic preferences. Rather, it depends only on the Prisoner's Dilemma-creating structural features of interaction.

maximization conception of rationality does not, by itself, bring in self-interestedness. As Gauthier puts it, such a conception of rationality introduces a purely formal, not a material, selfishness (*MbA*, p. 73).[12]

Gauthier, however, adds a third condition to the characterization of the bargaining agents of morals by agreement, namely, a type of self-interestedness. More precisely, he wishes to assume that their values are independent in another sense than that of agent relativity. Utility functions are to be defined independently of one another (*MbA*, p. 86). That is, the preferences of the rational bargainers do not range over the preferences of others. Gauthier thus assumes that individuals do not take an interest in the interests of others,[13] or weaker yet, they do not take an interest in the interests of those with whom they interact. The latter condition is that of Wicksteed's "nontuism," the former that of mutual unconcern.[14]

Gauthier believes that "this conception, of persons as taking no interest in one another's interests, is fundamental not only to economics, but also to moral theory" (*MbA*, p. 100). Mutual unconcern and nontuism here, we should note, are merely assumptions, albeit important ones. They are not meant literally to characterize humans.[15]

In morals by agreement, then, agents maximize the satisfaction of their mutually indifferent, or nontuist, (considered) preferences. Justice is to be identified with the principles to which such agents would agree, given the situation in which they find themselves.

Moral standing

To have *moral standing* is to be owed moral consideration. Depending on the particular account we adopt of moral consideration (e.g., natural duty theory, utilitarianism), depending on the particular moral virtue in question (e.g., justice, benevolence), or depending simply on the duties in question, moral standing admits of degrees. Thus, something that is owed more moral considerations than another may be understood

Gauthier, "Morality, Rational Choice, and Semantic Representation: A Reply to My Critics," *Social Philosophy and Policy* 5 (1988): 215.

[12] See also Gauthier, "The Incompleat Egoist," *The Tanner Lectures on Human Values*, Stanford University, May 10, 1983, p. 73.

[13] The phrase is from Rawls, *Theory of Justice*, p. 13.

[14] I simplify matters by focusing on mutual unconcern and on nontuism, but the story is actually more complicated than this. Some of the complications are discussed in my "The Relation between Self-Interest and Justice in Contractarian Ethics," *Social Philosophy and Policy* 5 (1988): 119–53, esp. 123–5. See also Peter Vallentyne, "Contractarianism and the Assumption of Mutual Unconcern" (chap. 5 in this volume).

[15] "Throughout our argument, non-tuism has served as an assumption . . ." (*MbA*, p. 329)

to possess greater moral standing. These matters of degree aside, there is an important distinction to be made between inclusion in the exclusion from the scope of (a) morality, and it is this distinction that is drawn by the concept of moral standing.

To make clearer the notion of moral standing, I introduce some additional notions. A *moral object*, I say, is something that is an object of moral consideration. A *direct* moral object is something to which (or to whom) that consideration is paid; an *indirect* moral object is something *about* or *concerning* which moral consideration is paid. The latter is a *beneficiary* of the moral consideration. Typically, direct moral objects are beneficiaries of moral considerations owed to them; thus they typically are also indirect moral objects. But this need not be. The different objects of moral duties can be determined by asking *to* whom/what and *regarding* whom/what are they owed. Suppose that Albert promises Beatrice to care for Calvin. Albert's duties would be owed to Beatrice *regarding* Calvin; the latter would be an indirect moral object of those duties. Were Albert not to care for Calvin in the requisite manner, he would fail in his duties toward Beatrice; though Calvin would fail to be benefited, he would not be wronged by Albert's delinquency.

In terms of this distinction, people typically are direct moral objects, or so we normally believe. Protected natural sites, national monuments, significant works of art might be examples of indirect moral objects. When we destroy the latter, we may be understood to fail in our duties to other people. Animists and others could use this distinction to give a different account of whom we fail to respect when we defile nature.

To have moral standing, then, is to be a direct moral object. Something that is merely an indirect moral object – for instance, a sculpture or a relic – would lack moral standing. It is important to note that the category of *moral value* is broader than that of moral standing. For an inanimate object such as a sculpture could have moral value without having moral standing, as it would in the case that it is a mere indirect moral object. Presumably, this is the account that most will want to offer of the moral status of the environment; Yosemite and the Grand Canyon would thus have moral value, without having moral standing.[16]

To have moral standing, then, is to be owed moral consideration, that is, to be a direct moral object.[17] It is important to note that moral standing

[16] The concept of a moral object is not to be confused with that of a moral *subject*. The latter is something that has moral duties or may be expected to give moral consideration to direct moral objects. Presumably, young infants can be direct moral objects without being moral subjects. Also, being a moral subject is not necessarily to have moral standing, for owing moral considerations to others does not entail that the latter owe one similar consideration.

[17] A worry about this characterization of moral standing, which is meant to be as neutral as possible between competing moral theories, is that it may not be applicable to utilitarian and other theories that do not make essential use of deontic notions. For, as I have explicated these notions, moral standing involves being *owed* certain considerations, being that to which (rather than *regarding* which) consideration is due. These deontic notions may introduce an individualist or agent-relative perspective and consequently bias the discussion against certain theories. At the same time, utilitarian theories do distinguish between those things that count morally, to which the principle of utility is applied, and those that do not. So my characterization of moral standing should be sufficiently neutral so as not to beg any important question.

is a *relation*: something has moral standing to the extent that something else owes it moral consideration. So something could not have moral standing if there were no other thing that owed it moral consideration. Further, and most importantly, we should note that it should not be supposed that moral standing must be *universal*. It would beg the question this essay addresses to suppose that all beings of a certain sort (e.g., rational agents) possess moral standing. But that is not the sort of universality that I have in mind. Rather, I want to note that we should not suppose that if some individual has moral standing, then it has standing in relation to all other individuals or moral subjects. The relations of moral standing may be particular (or even, in the extreme, pairwise). That is, it may be that Albert has moral standing in relation to Beatrice and Calvin but not in relation to Daphne; whereas Beatrice and Calvin owe Albert moral consideration, Daphne does not.[18] Moral standing, then, is a relation between classes of moral objects and moral subjects, and we may not suppose that membership in either of these classes is universal.

The question, then, is to determine what has moral standing (in relation to what). Or rather, the question is to determine how to determine what has moral standing (in relation to what).

Moral standing in rational-choice contractarian morality

In virtue of what, then, do entities have moral standing in morals by agreement?

One might think that contractarian justice accords moral standing to all and only agents who are members of the (hypothetical) "social contract." In this view, only agents capable of so contracting could acquire moral standing; though presumably not all of these will, in fact, acquire it (as I argue later). Participation in the social contract would thus be necessary for moral standing. But such a view depends on understanding contractarian ethics as involving contracting, whether actual or hypothetical, and that, I argue, is to take the metaphor of a social contract too literally.

[18] Indeed, this is precisely the implication of Gilbert Harman's relativist conventionalism. Peter Danielson suggests metaphors of firms and partnerships for our cooperative (i.e., moral) relations. The latter may be as complex as relations between employees, employers, partners, suppliers, business competitors, government agents, and foreign counterparts.

The metaphor of the social contract serves to express this particular conception of justice as an ideal convention, the terms of which may be thought of as the outcome of a rational bargain. The metaphor of the social contract is extremely misleading if it is taken, as it often is, in any other way. For agreement (or "contracting") here has only a heuristic value.[19] As I understand Gauthier's contractarianism, justice and related parts of morality are conventions, and the purpose of a hypothetical social "contract" is to determine the terms of the best (i.e., most rational) convention, at least for particular places and times. Hypothetical agreement determines maximal advantage.[20]

Why then is hypothetical agreement necessary for the determination of rational moral principles in morals by agreement? If the nonmoral world that constitutes the starting point for morals by agreement satisfies, for some set of individuals, the circumstances of justice, it is Pareto-inefficient. Suppose that in some such world there is only one outcome that is Pareto-superior and that it is in fact strongly Pareto-superior – that is, it is unanimously preferred to the status quo. Then this outcome would be uniquely mutually advantageous and would be selected by rational principles of distributive justice. Morals by agreement would not, for such a world, require hypothetical agreement to determine the content of principles of justice; mutual advantage would be both necessary and sufficient for such principles. This shows that hypothetical agreement is necessary only because of a feature of our world, namely, that there are many Pareto-superior outcomes and that we must choose among these. The function of hypothetical agreement, then, is to make this choice.

One may argue that contract or agreement has another role in morals by agreement. Its first role, it may be admitted, is that of a heuristic device to determine the content of acceptable moral principles. Its additional role may be to bind individuals to these principles. A contract of this type cannot *morally* bind contractors, for independently of agreement, there are no moral constraints.[21] How might such a contract bind?

[19] In conversation, the author of *Morals by Agreement* has expressed doubts about my claim that hypothetical agreement has *only* heuristic value.

[20] "Theories of hypothetical consent discuss not consent but cognitive agreement." Joseph Raz, *The Morality of Freedom* (Oxford: Clarendon Press, 1986), p. 81, n. 1.

Criticisms of contractarianism, to the effect that a hypothetical contract is no contract at all or that a nonmoral agreement does not bind, are consequently misplaced. For instance, see Ronald Dworkin, "The Original Position," in *Reading Rawls*, edited by Norman Daniels (New York: Basic Books, 1976), pp. 16–53.

[21] Note that I say that there can be no moral constraints *independently of* agreement. This is not to say that there cannot be moral constraints *prior to* agreement, for, in effect, Gauthier argues that there are some such constraints, namely, the proviso and the initial rights (*MbA*, chapter VII). "Although a part of morals by agreement, it [the proviso] is not a product of rational agreement. Rather, it is a condition that must be accepted by each person for such agreement to be possible" (*MbA*, p. 16).

Elsewhere, Gauthier "distinguish[es] contracts from other agreements by characterizing the former as exchanges of intentions to act that introduce incentives, whether internal or external, moral or other, to supplement or replace each party's motivation to attain the true objective of the agreement."[22] Contracts, so understood, are distinguished from purely coordinative agreements. While in both, all prefer the outcome of agreement to that of no agreement; in coordinative agreements, all prefer compliance with the agreement to noncompliance, given compliance by the others. Further, in coordinative agreements, each expects the others to comply, and intends to do so himself or herself. Contract, then, could play an additional role and thus would be instrumental in determining who has moral standing. Contract, as characterized before, might serve additionally to bind agents to the principles generated by agreement; thus, only agents who were members to the contract would have moral standing.

This does not follow, however. That is, the necessity of introducing contract to guarantee rational compliance does not require membership in a contract as a condition for moral standing. First, note that such a notion of contract is a substantive normative device, albeit not necessarily a moral one. Granting that such contracts introduce additional incentives, how are they available to individuals who find themselves in the circumstances of justice? Simply to postulate the availability of such contracts, without providing an account of their source, would be to beg the question. The problem is not that such contracts are moral and that their assumption would thus be question begging. Rather, the problem is that such contracts provide additional incentives (and consequently are normative), and an account of how they do this is needed. Now Gauthier does provide such an account when he tries to show that rational utility maximizers have a reason, in certain circumstances, to cease being "straightforward" maximizers and to become "constrained" maximizers.[23] His account, however, makes no essential reference to contract or agreement, or even to collective choice or action.

Requiring participation in a contract or agreement for moral standing is to understand contractarian theory as a type of consent theory. On

[22] "Hobbes's Social Contract," *Nous* 22 (1988): 73.

[23] *MbA*, chap. VI. Roughly, a straightforward maximizer chooses the utility-maximizing act at each decision node, whereas a constrained maximizer acts on a utility-maximizing policy, one which may require acts that are themselves nonmaximizing. Thus, for instance, in some one-play and finitely iterated Prisoners' Dilemmas, the former "defects" no matter what the others do, while the constrained maximizers "cooperate" when they believe that the others are similarly disposed. The success of Gauthier's argument is, of course, a subject of great controversy. Critical assessment of this part of Gauthier's theory is provided by the essays in Part III of this collection.

such an understanding, a particular act of *will* – participation in an agreement – is necessary for moral standing. Thus, someone who had not consented in the requisite manner, would lack moral standing. If the requisite act of will is taken to be participation in an actual agreement, then this understanding of morals by agreement as a consent theory is obviously mistaken. If, however, we understand consent differently, then there may be a sense in which morals by agreement is a consent theory. Joseph Raz suggests that:

> [c]onsent is given by any behavior (action or omission) undertaken in the belief that
>
> 1. it will change the normative situation of another;
> 2. it will do so because it is undertaken with such a belief;
> 3. it will be understood by its observers to be of this character.[24]

Rational-choice contractarianism does require, as a condition of possessing moral standing, a particular act of will that will constitute consent in Raz's sense. To this we now turn.

One of the conditions giving rise to the need for justice is the possibility of mutual benefit. Others are the capacity and willingness of rational beings to impose constraints on their behavior. In the absence of such conditions, it would appear, one has no reason to abide by the constraints of justice in one's conduct toward others.[25] This is important, for it effectively means that in the absence of (1) mutual benefit or of (2) the capacity or of (3) the willingness of others to be just, an individual is not constrained by justice in his or her behavior toward others.[26]

Supposing that the circumstances of justice be satisfied, morals by agreement understands rational humans, capable and willing to impose moral constraints on their conduct toward others, as moral subjects and direct moral objects (in relation to specified others). Thus, for this theory, as for most others, in normal circumstances, adult humans have moral obligations and are owed certain moral considerations. According to morals by agreement, then, some individual *A* has moral standing in relation to some person *B* if and only if

[24] Joseph Raz, *Morality of Freedom*, p. 81

[25] As Hume noted:

> Suppose, likewise, that it should be a virtuous man's fate to fall into the society of ruffians, remote from the protection of laws and government . . . 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, without concern for those who no longer merit his care and attention. (*An Enquiry Concerning the Principles of Morals*, sec. III, pt. I, p. 148)

[26] "This is plainly the situation of men, with regard to animals . . . ," ibid., p. 152.

(1) *A* and *B* are in the circumstances of justice,
(2) *A* and *B* are capable of imposing constraints on their behavior toward one another, and
(3) *A* so constrains his or her behavior toward *B*.

If and, it would seem, only if these three conditions are satisfied, then *B* owes *A* (some) moral considerations.

In this view, it sometimes may be the case that a rational agent lacks moral standing in relation to another. For instance, if some of the circumstances of justice are not satisfied (condition 1), then some agents will lack moral standing.[27] Or an individual – for instance, an infant – who lacks the capacities required for constraint (condition 2, which I call the agency requirement) may lack moral standing with regard to others. (I shall discuss exceptions presently.) Lastly, suppose that someone is unwilling to constrain his or her action toward another (condition 3). Then that individual lacks moral standing in his or her relations with others.

This last implication especially will strike many people as counterintuitive. Now ordinary moral intuitions, as Gauthier emphasizes, have no weight in fundamental contractarian moral theory (*MbA*, p. 269). However, note that this implication does not mean that one is permitted – or rather, not forbidden – from treating such creatures as one pleases. For they may still have moral value in the technical sense characterized earlier; that is, they may be – and presumably would be in the case of human infants – indirect moral objects. Thus, we would not be morally allowed to mistreat them.[28]

Still, infants and others about whom we care would lack moral standing on this view. And leaving aside the matter of the counterintuitive nature of this implication, consider the more serious problem of the possible incoherence of such implications and the conception of value that Gauthier invokes.[29]

Consider the case of Emil, Frederica, and Gerhardt. Emil, although rational, exploits Frederica. He does this because he is able to do so and it is to his advantage. Emil and Frederica do not find themselves in the circumstances of justice, perhaps due to the former's superior strength.[30]

[27] See *MbA*, p. 17.

[28] This essentially is the strategy Mary Ann Warren takes, perhaps unwittingly, to avoid justifying infanticide. "The needless destruction of a viable infant inevitably deprives some person or persons of a source of great pleasure and satisfaction. . . ." "Postscript on Infanticide" appended to "On the Moral and Legal Status of Abortion," in *The Problem of Abortion*, 2nd ed., edited by Joel Feinberg (Belmont, CA: Wadsworth, 1984), p. 117.

[29] What follows is a modification of a case I discuss in "The Relation between Self-Interest and Justice in Contractarian Ethics," pp. 146–8.

[30] This case is explicitly considered by Hume:

> Were there a species of creatures intermingled with men, which, though rational, were possessed of such inferior strength, both of body and mind, that they were

While Gerhardt does not find himself in the circumstances of justice with regard to Frederica, he does with regard to Emil. According to morals by agreement, poor Frederica stands outside of the protection of justice. She finds herself in a Hobbesian state of nature, where she and others are at liberty to do as they please with one another. Emil and Gerhardt, however, are in a different situation. They are morally bound to one another insofar as rational cooperation is mutually advantageous.

While Emil is a purely self-interested fellow, Gerhardt is not. Indeed, the latter is most upset by the former's treatment of Frederica. Gerhardt does not consider Frederica's virtual slavery to be unjust, for it is neither just nor unjust according to morals by agreement. It is rather that he simply takes an interest in her interests. He would like to liberate Frederica from her plight, by force if necessary. However, morals by agreement does not permit him to do so. For Gerhardt is morally obliged to respect Emil's life, liberty, and possessions, as well as to accord him the distribution of the social surplus afforded him by the principle of minimax relative concession.

Suppose Gerhardt were to consider himself in a Hobbesian state of nature with regard to Emil and thus be able to liberate Frederica? This would be irrational, according to Gauthier's account of morals by agreement, for he is not in such a state of nature given his self-interested preferences, that is, given the assumption of mutual unconcern. Indeed, it would actually be unjust for Gerhardt to come to Frederica's aid!

Note that the objection here is not the standard sort of criticism made of morals by agreement, that it violates one of our intuitive moral judgments. I do not deny that this is the case; the implications are morally *unintuitive. But that is not the objection. (Nor is it an objection, given* the rational-choice methodology.) Rather it is that rational, other-regarding individuals, with utility functions like ours, are not moved by justice in cases such as these. The problem is not (merely) one of compliance; it is not that we would refuse to comply with norms that we would otherwise endorse. It is that we would find the norms themselves unacceptable in such situations.

Does morals by agreement have such implications? Recall the two motivational assumptions used by Gauthier to characterize the rational bargainers of morals by agreement: Rawls' mutual disinterest or unconcern and Wicksteed's nontuism. This case presupposes the first form of self-interestedness. If parties are disinterested in the manner postulated

> incapable of all resistance, and could never, upon the highest provocation, make us feel the effects of their resentment; the necessary consequence, I think, is *that we should not, properly speaking, lie under any restraint of justice with* regard to them, nor could they possess any right or property. . . . (*An Enquiry Concerning the Principles of Morals*, sec. III, pt. I, p. 152)

by Rawls in his theory of justice, then the previous objection goes through. If, however, the type of self-interestedness assumed is that of nontuism – taking no interest in the interests of those with whom one interacts – then the objection, it may be argued, fails.[31] For the assumption of nontuism does not rule out Gerhardt's concern for Frederica in the determination of the principles governing the relations between Gerhardt and Emil.

Suppose that Gerhardt's tuistic preferences are such that in the hypothetical bargaining situation, he would insist that Frederica be given moral standing. He cares about her to such a degree that it would not be rational for him to interact with others except on the condition that she be accorded moral standing. An *intermediate* case would be one where Frederica would be given the status of an indirect moral object. But such a case is not theoretically interesting. More important is the previous case, where Gerhardt's preferences are such that Frederica would have to be given the status of a direct moral object.

Were agents to care in this manner for (some) others and were morals by agreement to take this into account, then there would be a second way in which individuals could acquire moral standing – namely, by being the object of the preferences of an agent who finds himself or herself in the circumstances of justice. Thus, Frederica might acquire moral standing in her relations with Emil through being the object of Gerhardt's preferences. And, similarly, children and others who do not meet the conditions for agency would presumably be accorded moral standing.

In addition, then, to the straightforward way in which agents can *acquire moral standing*, discussed earlier, there is a second, indirect manner to acquire such standing. An individual who is not an agent or who is not in the circumstances of justice can acquire moral standing by being the object of the preferences of others. We distinguish, then, the two ways in which individuals can acquire moral standing in morals by agreement:

> *Primary moral standing: A* has moral standing in relation to *B* if (1) *A* and *B* are in the circumstances of justice, (2) *A* and *B* are capable of imposing constraints on their behavior toward

[31] Gauthier so argues in response to my original case:

> In terms of his interaction with Adolf, Charles's concern with Bécassine is nontuistic. In any interaction, concerns with the interests of a third party are nontuistic. . . . Nontuism thus does not have the implications for contractarian moral theory that the assumption that all preferences are self-interested or self-directed would have.

Gauthier, "Morality," p. 215. In the text of *Morals by Agreement*, however, it is not always clear which motivational assumption – mutual unconcern or nontuism – is being made.

one another, and (3) *A* so constrains his or her behavior toward *B*.

Secondary moral standing: *A* has moral standing in relation to *B* if (1) *B* and *C* are in the circumstances of justice, (2) *B* and *C* are capable of constrained action, (3) *C* constrains his or her acts toward *B*, and (4) *A* is the object of *C*'s preferences, that is, *C* cares sufficiently about *A* that it would not be rational for *C* to cooperate with *B* unless *A* were accorded moral standing in his or her relations with *B*.

"Secondary moral standing," it should be emphasized, is merely a manner in which moral standing can be acquired. Someone who acquires moral standing in this way has genuine moral standing. It is secondary only in the sense that were no one to have primary moral standing, no one could have secondary moral standing.[32]

Morals by agreement, then, may accord moral standing to infants, the infirm, and others in the second way indicated before. Presumably, even nonhuman animals may be accorded some moral standing in this manner.[33] What creatures are recognized as direct moral objects depend heavily on the particular other-regarding preferences of the nontuistic agents of morals by agreement.[34]

Self-interest and moral standing

We have determined two ways in which something can acquire moral standing according to morals by agreement. It is difficult to see how there might be other ways. What I wish now to explore is the manner in which the particular implications of morals by agreement, regarding

[32] In *Persons, Rights, and the Moral Community* (New York: Oxford University Press, 1987), pp. 152–3, Loren E. Lomasky argues that

> all those who are characterized by property F (project pursuit) have property G (possession of basic rights). . . . That there must be F's in order for there to be G's does not entail that *only* F's are G's. Some beings who lack F can yet be G's by piggybacking on those who are F's.

[33] Whether nonsentient things could acquire moral standing in this secondary way would depend on whether the notion of owing moral considerations requires that the object have interests. One would suppose that it would, but insufficient content has been given to the notion to permit an exploration of this question. Some of my remarks about justice and benevolence that follow have bearing on this question.

[34] Wayne Sumner has argued that the range of beings accorded moral standing by contractarian theory is determined largely by the range of the concerns of the agents. I failed to understand his point at the time, but it now seems correct to me. See Morris, "Value Subjectivism, Individualism, and Moral Standing: A Reply to Sumner," and Sumner, "A Response to Morris," in *Values and Moral Standing, Bowling Green Studies in Applied Philosophy* VIII, edited by Wayne Sumner, Donald Callen, and Thomas Attig (Bowling Green, OH: Bowling Green State University, 1987), pp. 16–21, 22–3.

the moral standing of individuals, are dependent on particular assumptions about preference and value. Gauthier assumes that value is determined by (coherent and considered) preference and that the latter, for the purposes of fundamental moral theory, are nontuistic. What if we drop either of these two assumptions? I explore the matter of nontuism, leaving to another occasion that of subjective value.[35]

Recall the interactions between Emil, Frederica, and Gerhardt. Suppose that the relevant assumption is that of mutual unconcern. Then Frederica lacks moral standing. If we replace this assumption with nontuism, then the conclusion no longer follows; should Gerhardt care sufficiently for Frederica, then she will acquire moral standing in the second, indirect manner.

Let us distinguish between several motivational assumptions that appear in *Morals by Agreement*. I enumerate them in order of weakness.[36] *Nontuism* requires that agents not take an interest in the preferences of those with whom they interact. *Mutual unconcern*, Rawls' mutual disinterest, requires that agents not take an interest in the preferences of others. *Egoism* I characterized earlier as requiring that all of an agent's preferences be self-regarding. Egoism, thus characterized, conflates rationality and prudence. Egoism is stronger than mutual unconcern; mutually disinterested agents need not be egoists as they may take nonagents or nonhumans to be the object of their concern.

Further, there is the general condition of asocial motivation or *asociality*, which Gauthier introduces late in the book (*MbA*, p. 311), requiring that descriptions of the preferences of individuals not make mention of other individuals. This condition is stronger yet than egoism, as asocial agents cannot have the concerns with, for example, relative status possible for egoists. Lastly, there is what I shall call the assumption of *private consumerism*, for want of a better label. This is the motivational assumption, often made in the theory of perfectly competitive markets, that requires that agents' utilities are functions only of "commodities," none of which are public goods.[37] We might make more determinate the character of "economic man" by characterizing his preferences in this way.[38] The assumption of private consumerism is thus the assumption that agents are "economic men."

[35] In "Agent-Relative Value, Justice, and the Compliance Problem" (manuscript), I argue that Gauthier's subjectivist account of value is stronger than is required by most of his substantive conclusions and that these would not be affected were it to be replaced by any of several nonsubjectivist accounts.

[36] Where by "weakness" I mean that a strong assumption entails a weaker one, but not vice versa.

[37] That is, commodities are private goods, divisible and excludable.

[38] This would accord with much of what Gauthier says about economic man in Chapters X–XI of *MbA*, as well as in his earlier "The Social Contract as Ideology," *Philosophy and Public Affairs* 6 (1977): 130–64. It would not, however, be consistent with his character-

I have distinguished, in order of weakness, various motivational assumptions that are invoked or otherwise mentioned in *Morals by Agreement*: nontuism, mutual unconcern, egoism, asociality, and private consumerism. In the case of Emil et al., only the condition of nontuism allows Frederica (secondary) moral standing.

Weaker yet than nontuism is simply to allow preferences to range over the interests of anyone, whether one is interacting with them or not.[39] To do this, however, allows the possibility of "double counting." Suppose that

> I, considering us equally fond of cake, prefer that each of us get half, not only to your having a larger share but also to my having it, and if you prefer more cake for yourself to less, whatever I get, then it seems implausible to suppose that a rational and fair division gives you three-quarters of the cake and me one-quarter.[40]

Your preferences in this example are counted twice, that is, weighted more heavily than mine. Double counting, at least in the formulation of principle of distributive justice, is thought by many to be counterintuitive.[41]

Double counting might be inefficient, however, and rational agents may find it mutually agreeable to bargain without referring to their tuistic preferences. Or so Gauthier argues in a recent defense of nontuism:

> As I now see it, social institutions and practices should be justified by an appeal to a hypothetical agreement based largely on the nontuistic preferences of the parties concerned, because each person expects ex ante to benefit if she forgoes the inclusion of her tuistic preferences in determining social arrangements provided others do the same. Double-counting will be ruled out, as, of course, will preferences directed at the frustration of other's [sic] preferences. . . . Positive, but not negative, regard for others will then be furthered.[42]

ization of "economic rationality" in "Economic Rationality and Moral Constraints," *Midwest Studies in Philosophy* 3 (1978): 75–96.

[39] This is the strategy of Loren Lomasky in his conventionalist theory of rights, where he assumes that agents come to take an interest in the interests of those with whom they interact. See Lomasky, *Persons*, p. 65ff.

[40] Gauthier, "Morality," p. 214. See also Morris, "The Relation," pp. 137ff.

[41] See the reviews by Gregory S. Kavka, *Mind* XVCI (January 1987): 117–21, and by Loren Lomasky, *Critical Review* 2 (Spring/Summer 1988): 36–49. Kavka and Lomasky, however, are mistaken in attributing these counterintuitive implications to Gauthier's morals by agreement, as they are blocked by the assumption of nontuism.

[42] "Given this revision [from the account offered in *Morals by Agreement*], the exclusion of tuism from the justification of social institutions and practices and so from the public realm rests simply on an empirical fact, if, as I suppose, it is one, about the role that tuistic and nontuistic preferences play in our concerns." Gauthier, "Morality," p. 216.

Robert Goodin similarly argues that social-choice theory can exclude certain preferences without having recourse to nonutility information; thus, social-choice theory need not be "welfarist." See his essay, "Laundering Preferences," in *Foundations of Social Choice Theory*, edited by Jon Elster and Aanund Hylland (Cambridge: Cambridge University Press, 1986), pp. 109–21.

I shall not pursue the matter of a plausible defense of nontuism. Instead, let us return to the suggestion that we not restrict the range of concerns to be considered and that we allow preferences to range over the interests of anyone, whether one is interacting with them or not. To do this would be to be make an assumption yet weaker than nontuism. What implications would this have for questions of moral standing?

The implications are interesting, although explaining them is a rather complicated matter. Suppose that Henrietta and Ivan are interacting in a situation that meets some but not all of the circumstances of justice. While both are rational agents, Ivan is weak, and Henrietta is sufficiently strong that she can gain more from coercive interaction with Ivan than from cooperation. Assuming nontuism, or any of the stronger assumptions, it follows that Henrietta is not bound by justice toward Ivan – unless, what I am assuming not to be the case, that Ivan has secondary moral standing.

The fact that Henrietta is *not obligated* by justice to accord certain treatment to Ivan – for example, to respect his life, liberty, and possessions – does not, of course, entail that she is *forbidden* from doing so. Morals by agreement determines what requirements morality imposes on Henrietta; she clearly is permitted to do that which she is not required to do. Still, she is not required, by morals by agreement, to respect Ivan's moral rights – for he has none (in his relations with Henrietta). The assumption of nontuism prevents Ivan's acquisition of moral standing.

Suppose that Henrietta refrains from exploiting and otherwise harming Ivan because the satisfaction of his preferences are one of the values of her utility function. I phrase the matter in a technical way, so as to leave open, for the moment, the nature of Henrietta's concern. Let me now introduce some distinctions that do not appear in *Morals by Agreement*.

It is traditional to distinguish the moral virtue of justice from those of friendship, courage, moderation, and the like. These virtues are not mentioned in the pages of *Morals by Agreement*.[43] Nor is that of benevolence. Henrietta may care about Ivan, and for that reason refrain from harming him. But she may also be moved by considerations of benevolence, that is, by *moral* considerations over and above her (nonmoral) sympathies for him. Can morals by agreement make sense of, that is, generate obligations of benevolence? The answer is unclear, at least to me. We may interpret Hume as a contractarian about justice (and prop-

[43] Modern moral theories, especially social theories such as contractarianism, tend to classify such virtues as largely nonmoral. And David Gauthier suggested, after a lecture at UCLA in 1983, that he would relegate these to the domain of psychology.

erty)[44] and note that he counts benevolence as a moral virtue, and an important one at that. Hume's account of benevolence, however, is distinctively not contractarian or even conventionalist. Rather, he offers a moral sense account.[45] So his approach is foreign to that of Gauthier's morals by agreement.[46]

Let us distinguish between being the (direct) object of considerations of justice and of considerations of benevolence. We could say that a being that is a direct moral object both of considerations of justice and of considerations of benevolence has *full moral standing*. Something that is a direct moral object only of justice or only of benevolence (but not both) has *partial moral standing*. Something that is neither the proper direct object of considerations of justice nor of benevolence is said to have *no moral standing*.[47] Thus, mere indirect moral objects, although the (indirect) object of moral consideration, will lack moral standing. The American flag, Yosemite Park, the Louvre, for instance, will be protected by morality by being indirect moral objects of our duties to each other; they will lack moral standing, however, since *they* are neither owed considerations of justice nor of benevolence. Hume, then, generates partial moral standing – being the direct object of considerations of benevolence – by reference to sentiments widely, if not universally, possessed by humans; what I have called full moral standing requires his more complete, contractarian account.

Suppose, then, that we drop nontuism and related assumptions. Morals by agreement might then accord Ivan partial moral standing in his relations with Henrietta; the former, it might be argued, is a suitable object of the latter's benevolence. A more interesting possibility is that of extending full moral standing to him. Recall that Ivan is an agent, that is, that he is capable of acting intentionally and imposing constraints on his behavior. He is, I am supposing, also willing to do so. His problem, however, is that he does not find himself in the circumstances of justice in his relations with Henrietta, given her superior strength. Now if we drop the assumption of nontuism and factor in the latter's concern for Ivan, then it is possible that the circumstances of justice will be satisfied. That is, it is possible that Ivan and Henrietta might be in the circumstances of justice *given the latter's other-regarding preferences*, an effect of the "double counting" of Ivan's preferences.[48]

This manner of extending full moral standing may not, however, work for those incapable of cooperating, that is, for nonagents. Thus, infants do not acquire moral standing in this way.[49] Their moral standing has to be secondary. We should note, however, that typically it is the moral standing of infants *in relation to nonparents* that is secondary. It is not clear how it is that infants acquire moral standing *in relation to their parents*. Consider the case of Johann, the son of Katherine and Luigi. Suppose that Johann's parents are the only people (with primary moral standing) who care about him, or at least care about him sufficiently for him to be accorded secondary moral standing in his relations with others. Then, in light of the conclusions drawn before, it is not clear how Johann can acquire moral standing in his relations with his parents. Presumably such a case, where the only people (with primary standing) who care about an infant are his or her parents, is rare. Still, it is an interesting (and bizarre) implication that such infants lack moral standing in their relations to their parents, though not necessarily in their relations to others. It appears that someone need not have moral standing in relation to the agent(s) who is the vehicle for one's secondary moral standing.[50]

I have explored the complicated manner in which morals by agreement accords individuals moral standing. I have also discussed the ways in which some of the substantive implications of Gauthier's theory depend on certain motivational assumptions that he makes. Some of these implications may be avoided by weakening or otherwise altering these assumptions. This fact may not fully satisfy critics, but it advances nonetheless our understanding and appreciation of the theory.

[44] See Gauthier, "David Hume, Contractarian," *Philosophical Review* 88 (1979): 3–38. See also J. L. Mackie, *Hume's Moral Theory* (London: Routledge and Kegan Paul, 1980), chap. VI, and Rawls, *Theory of Justice*, pp. 32–3.

[45] Hume, *An Enquiry Concerning the Principles of Morals*, sec. II

[46] In addition, note Gauthier's antisentimentalism

> Hume believed the source of morality to lie in the sympathetic transmission of our feelings from one person to another. But Kant, rightly, insisted that morality cannot depend on such particular psychological phenomena, however benevolent and humane their effect, and however universally they may be found. (*MbA*, p. 103. See also pp. 309, 326–9, 338–9.)

[47] This crude, tripartite classification of degrees of moral standing assumes that justice and benevolence are the only moral considerations that can be owed to entities. Note that this assumption restricts moral standing to entities that can be owed considerations of justice or benevolence, which presumably will be creatures with a welfare. See footnote 32.

[48] Additionally, it may also be, as is the case with many human relations, that the benefits to Henrietta from interacting with Ivan can be obtained only noncoercively. And the relation that Henrietta desires to have with Ivan may be one that presupposes the minimal equality of justice

[49] The moral status of infants, it seems to me, is determined in part by their *potential* agency. I do not know how potential agency determines standing, so what follows ignores this aspect of the matter. The considerations mentioned in footnote 47 are also important in this regard.
Suppose that Luigi becomes abusive of Johann and is opposed by Katherine. Given the latter's concern for Johann and given that Luigi and she remain in the circumstances of justice, then Johann may acquire secondary moral standing in his relations with his father.

Part II

Minimax relative concession and the Lockean Proviso

Overview of the essays

Peter Danielson argues that Gauthier is mistaken to take the initial bargaining position to be that of noncooperation constrained by the proviso. For rational bargainers will only be concerned with improving upon their prebargain positions. Since their prebargain positions are *not* based on any constraint from the proviso, Gauthier is mistaken in his claim that rational bargains must be based on the proviso. Furthermore, Danielson argues that rational agents would treat property rights as one of the issues that will be settled by the bargain – and not as something that is built into their initial bargaining positions.

Don Hubin and Mark Lambeth argue that Gauthier's use of the proviso is inappropriate as a foundation for moral rights on the following grounds: (1) The proviso permits one to worsen the situation of others very significantly, when doing so is necessary to prevent one's own situation from being worsened only slightly. (2) The proviso permits one to kill, beat, or rob others when someone else would do so if one didn't. (3) The proviso permits one to use people in all sorts of horrible ways as long as one also helps them in various ways so that the *net effect* on them is positive. (4) Because worsening is understood in terms of people's subjective preferences, the proviso can (under appropriate circumstances) prohibit all sorts of activities (such as going for a walk) simply because someone else prefers that one doesn't engage in that activity.

In his essay, Jan Narveson starts by considering Gauthier's minimax relative concession-bargaining solution. After briefly questioning one of the assumptions on which Gauthier's solutions rests, Narveson argues, against Gauthier, that bargaining theory makes sense only in a context

in which the parties already have rights. And if that is so, then the (moral?, rational?) requirement to respect such rights does not depend on conformity with any bargaining solution. Narveson then goes on to discuss Gauthier's use of the Lockean Proviso. Although he agrees with Gauthier that the appropriate initial bargaining position is the hypothetical outcome of noncooperation constrained by the Lockean Proviso, he argues that the proviso is relevant only for determining the distribution of the benefits of cooperation (after agreement) – not for (re)distributing the benefits of coercion that took place prior to agreement.

Jean Hampton criticizes Gauthier's bargaining solution by arguing that it is more plausible that rational agreement would be based on a principle that allocates benefits in proportion to contribution to the social surplus (and not – as Gauthier claims – so as to minimize the maximum relative concession).

Wulf Gaertner and Marlies Klemisch-Ahlert discuss Gauthier's minimax relative concession-bargaining solution. They start by discussing Nash's solution and the Kalai–Smordinsky solution. They finish by giving an axiomatization of Gauthier's solution (the first that has ever been given) that clearly isolates the differences between his solution and the others.

A caveat: Gauthier defends the view that the *rational* permissibility of strategic choices is determined by the application of the minimax relative concession principle to the Lockean-Proviso-constrained noncooperative outcome. He also holds that a strategic choice is *morally* permissible if and only if it is rationally permissible. Consequently, he holds that the *moral* permissibility of strategic choices is also determined by the application of the minimax relative concession principle to the Lockean-Proviso-constrained noncooperative outcome. There are two issues here: (1) Are minimax relative concession and/or the proviso relevant for rational choice? (2) Are they relevant for moral choice? In Gauthier's contractarian theory, these two issues are coextensive, but conceptually they are distinct. Since authors do not always clearly distinguish these two issues, readers should be careful to determine exactly which claim is being assessed.

7. The Lockean Proviso*

Peter Danielson

There are numerous ways to divide the fruits of beneficial social cooperation, which satisfy differentially the competing interests of the would-be cooperators. Selecting one of these principles of distributive justice is what we will call the contract problem. In spite of sharp disagreement over the solution of this problem, there is wide agreement over how to picture it, so we focus on a diagram (see Fig. 7-1).

In this simple example, the problem is to divide the social product between two groups, the masters and the slaves. The utility of a distribution to the masters is represented along the horizontal axis; the utility to the slaves along the vertical axis. There are two important sets of points in this space. First, the set of possible optimal distributions ranges from almost all to the slaves in the upper left to almost all to the masters in the lower right. (It is concave to the origin, reflecting the greater utility of equal distributions.) Second, there is a set of alternatives to full social cooperation. Since they fall short of full cooperation, these states of nature (labeled with I's for initial positions) are clustered in the southwest. For example, there is the natural distribution, I_N, where the masters coerce the slaves. In contrast, there is the noncoercive I_B, where only what we call personal rights are respected. Note that here the slaves do better and the masters worse than at I_N. There is also the Lockean state of nature, I_C, where property rights are respected as well as rights against coercion. I_C is northeast of I_N and I_B as everyone does better here, although, as we shall see, the masters do better than the (ex)slaves.

*Excerpted by permission of the *Canadian Journal of Philosophy* from "The Visible Hand of Morality," by Peter Danielson, in *Canadian Journal of Philosophy* 18 (1988): 357–84.

only if he compensates the other. But, absent our additional assumption, Gauthier does not believe that justice requires any such compensation. He says:

> If . . . [the] 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.[21]

It seems implausible that desire for positional advantage should alter the situation so drastically in the Robinson Crusoes case.

Now, of course, this case is just a variant of the McDonald/McDougal example. But perhaps the variation makes clear what is at issue. At least for many, the conviction that those with the positional advantages ought *as a matter of justice* to compensate those with less seems to depend on the existence of social relations. McDonald and McDougal were individuals set in a situation in which there typically is social interaction. Although we did not assume any such interaction, still one's convictions about the case may be influenced by the setting. In the Robinson Crusoes example, such extraneous influences are not present.

The situation is worse if we remind ourselves that in order to avoid the problems with Dr. Demento, Gauthier must be concerned not with an action's effect on people, but with its effect on people's utility. Given this, we can imagine that the two Robinson Crusoes lack the ability to communicate with each other and neither ever learns of the existence of the other. Let us suppose that their desires remain the same (i.e., that each would prefer to be better off than someone else to being the lone person in existence and that each would prefer this latter situation to being worse off than someone else). It turns out that for the energetic Crusoe to use his abilities to the fullest is a violation of the proviso (even in their ignorance of one another's existence) – an excusable violation, no doubt, but a violation just the same. This seems wrong.

* * *

What these cases suggest is that the problem in Gauthier's theory is not with the attempt to describe the initial situation for a contractarian theory of justice in terms of a state of nature in which there have been no rights violations. Indeed, although others have criticized this approach, we find it to be a major attraction of his theory.[22] The problem is that even Gauthier's version of the proviso is not an adequate foundation for rights.[23]

[21] Ibid., p. 221.

[22] See, for example, "The Lockean Proviso" by Peter Danielson and "Gauthier on Distributive Justice and the Natural Baseline" by Jan Narveson, both in this volume, Chapters 7 and 9, respectively.

[23] We are indebted to Daniel Farrell, David Gauthier, and an anonymous referee for *Dialogue* for providing helpful comments on an earlier draft of this paper.

9. Gauthier on distributive justice and the natural baseline

Jan Narveson

Introduction: Gauthier's contractarianism

Contractarians hold that the fundamental principles of morals are the objects of something very like an "agreement," which in turn is the outcome of what is in some sense a "bargain." Just what sort of an agreement, "made" how and in what circumstances, is a matter on which different theorists in the contractarian tradition have given very differing accounts.

Gauthier's *Morals by Agreement* is the latest, the most sophisticated, and, in my view, by far the most compelling and perceptive effort in this great tradition. Let me briefly summarize what I take to be the distinctive points in his account. Gauthier sides with Hobbes and against Rawls, for instance, on the amount of idealization that goes into the construction of the appropriate "starting point" of the Social Contract. We do not reason from behind Rawls' Veil of Ignorance, but instead in full view of our assorted individual characteristics – but also, of course, in full view of our fellows. And, as with Hobbes, we begin with no moral presuppositions. Morality is to be the set of interpersonally applicable rules on which reason, driven by our actual (though considered) preferences, tells us to agree, and it tells us to agree on them despite the lack of any assumption of fellow feeling, love, or charity, or even of such philosophically popular constraints as universalizability, impartiality, or equality. We accept only such of these as can be seen to fall out of our fundamental project, which is to make the best life we can for ourselves in view of our actual situations vis-à-vis our fellows who, like us, are also rational pursuers of interests. Lacking moral constraints, we do worse; possessing them, we do better. Gauthier, however, also eschews the Hobbesian Sovereign. Rational people will internalize an

assortment of constraints on their pursuit of advantage, even though they reason from advantage in adopting the constraints. Yet once adopted, rational individuals adhere to them.

The theory comprises three subtheories: (1) a theory of bargaining, minimax relative concession (MRC), (2) a theory of compliance, constrained maximization, and (3) a theory of the appropriate natural baseline for the social contract, (his version of) the Lockean Proviso on acquisition. In this inquiry, I query the first and third of these in some respects. These are queries rather than major dissents, both in the sense that I am in very broad agreement with Gauthier on most matters, and in the sense that I have considerably less than total confidence in both of the modest dissents put forward here. Both, however, could make some, perhaps appreciable, difference to the practical implications of his theory.

My queries are motivated by the concerns implied in the preceding thumbnail sketch. Gauthier's project is to supply fully nonquestion-begging foundations for what are recognizably moral outcomes. Thus, his version cannot require the participants to take up unreal positions, positions they could not occupy in real life. Nor can it assume that the participants are impartial or natively equipped with respect for their fellows or for reason, either in some disputable version or even in the thin and hopefully less controversial version employed in the theory. If we are to emerge with a morality that is impartial, as Gauthier insists, then a rationale for this impartiality must be found. The rationale that the results agree with our pretheoretical intuitions is rejected: we wish to convince the previously unconverted, not just those previously disposed to agree.

As I see it, to motivate this impartiality, we need merely remind ourselves that we seek intersubjectively valid rules, rather than arbitrary personal edicts. Moral constraints are administered, fundamentally, from within. They must have the support of all reasonable persons – there is no one else but ourselves to do the job. But unless all support them, the support of any one individual becomes less rational. Now, we assume that each person naturally appraises situations from the perspective of her own values, which are likely to be highly biased: she is anything but impartial by nature. Why, then, must the rules we agree on be unbiased in relation to all actual participants? Because if they were not, then some would be called on to sign into an agreement that is either *suboptimal* – she could do better with no one doing worse – or *unfair*. In either case, Gauthier supposes, those individuals would have reason to refrain from accepting the proposed rules. But lacking their support, we would find ourselves on the skids back to the horrors (a.k.a. "suboptimalities") of the State of Nature. Self-interested choosers, if they know what they're doing, choose rules for the containment of

externalities because this yields a greater benefit for each, provided that all comply, than they would be able to attain in the absence of such rules. Similarly, they choose impartial rules for the distribution of *cooperative* benefits because each wants as much as possible and has no reason to settle for less than an equal share.

That suboptimality would yield the desired adherance is, in my view, plausible, and is not questioned here. Unfairness, however, brings up other and trickier issues. "Unfair" rules may be said to disproportionately reflect the distribution of relevant claims. When what we are negotiating for is distributive shares, then the outcome must reflect our prior claims, if we have any. And here we enter the vexed area of assertions concerning "equality." One way for the Social Contract to be unfair would be for it to provide unequal benefits for some when the prior claims of all are equal. Another would be for it to provide equal benefits when claims are in fact unequal. We must steer between these opposite errors.

This brings us to the two problems in Gauthier's theory that I wish to focus on in the remainder of this essay: (1) the status of his proposed principle of distributive justice, minimax relative concession, and (2) the status of his Lockean Proviso, specifically as it related to the status of preagreement gains in the postagreement situation. We discuss each in turn.

Gauthier on *distributive justice*

In Gauthier's view, 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" (p. 113).[1] He introduces his discussion of bargaining by distinguishing two domains of justice: on the one hand, it prohibits taking advantage, gaining at others' expense. Here the rules avoid "mutually destructive conflict" (p. 115). On the other, it is concerned with "the cooperative provision of mutual benefits" (p. 114) made possible by the fact of variable supply, which can be positively affected by cooperation. In cooperative production, there will be a cooperative surplus, not available without the participation of each cooperator. The question is how these surplus goods are to be divided among those participants.

The general bargaining problem is the problem of what principle of distribution is rational for such cases. Gauthier's project is to arrive at the appropriate principles, calling for the right constraints on the part of those concerned. These principles, setting the constraints required

[1] David P. Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986). Page references to this book are made in unattributed parentheses in the remainder of this essay. All other references are footnoted.

for rational cooperation, are arrived at *through* bargaining, but "are no part of the bargaining process" (p. 129). The bargainer is simply trying to maximize his utility: "each person's behaviour must be a utility-maximizing response to her expectation of others' behaviour. . . ." (p. 129).

How to proceed? To begin with, it is axiomatic that the outcome of the bargain for each bargainer, what she carries *from* the table, must, if bargaining is to be rational, be an improvement for her relative to the status quo: she must leave with more than she had when she came.[2] (Call this "condition A.") Each person then puts in an initial "*claim*" for some portion of the surplus in question. This corresponds to the opening round in haggling between buyer and seller in, say, an Arab market. How big is the opening claim? Since the first condition is satisfied by any distribution that leaves everyone at all better off, the prima facie answer is supplied in the first of Gauthier's four "conditions on rational bargaining," (i) that one claims *all*, or more precisely, just short of all, of this surplus (subject to one further restriction: what one claims is only the nearly-all of "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 cooperative interaction" (p. 134). For, obviously, one is otherwise taking advantage, by free-riding on those who did the producing.) Then the bargaining problem is the problem of where, rationally, to settle among the infinitude of possible but mutually incompatible distributions determined by each person's maximal claim, giving to each of the others the minimum that would make it rational for them to participate at all; and so on through all the other distributions, each giving a portion to each participant that continues to satisfy condition A. Here enters Gauthier's ingenious and attractive solution.

We identify the cooperative surplus itself by finding, first, what each person could get without participating. As seen above, condition A requires giving that person at least as much as his or her next-best available option from outside. This, therefore, can be subtracted from the surplus available for bargaining. Next, we must be able to characterize each person's claim in a way that makes that claim *comparable* to the claims of others. The utilitarian idea of cardinal, interpersonally comparable utility is rejected as the appropriate unit, for good reasons that we can't go further into here (pp. 126–9). Gauthier chooses as the appropriate measure that *relative fraction* of the bundle of goods which is the potential surplus in question going to each person. We assume that each wants as much as possible of that surplus – it might be money, for instance – and that each would find it rational to participate for any part of that surplus; fractions of it, then, make a plausible

[2] Our second problem, discussed in what follows, concerns one aspect of this requirement.

choice for comparison: ". . . the rationale turns on an interpersonal comparison of the proportion of each person's potential gain that he must concede . . . ," (p. 139) – remembering, of course, that what each person gains, above his minimum, is an exact inverse of what he loses, below his maximum, through concession – the gainer of 350 out of a potential 500 has foregone 150, whereas her counterpart, who has gained only 35, has foregone only 15 (pp. 138–9).

The argument then proceeds as follows. For each person, (ii) there must *be* a point at which he would be prepared to settle if need be; for otherwise, there can be no distribution of the surplus. (We can envisage the problem facing a set of potential cooperators deciding in advance on how to divide the gains, so that if there is no conclusion, then there is also no cooperation and hence no surplus to divide.) Each bargainer, being rational, knows this about each other bargainer. Then comes the crucial move. In Gauthier's words, "(iii) . . . Each . . . 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)" (p. 143). But being rational, (iv) he will therefore settle for nothing *less* than that either. For that point is by definition one that it is possible for all to settle at, so it is feasible; yet anything less would be irrational, since one doesn't have to settle below it.

Interesting, and plausible. But we must ask what makes condition (iii) plausible? What Gauthier says is that it "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." (pp. 143–4). This sounds good. But there is a problem. The argument talks of the "equal rationality" of the bargainers. How is that to be understood? Gauthier's theory of rationality is the maximizing theory (cum Constrained). Now presumably either one is a maximizer or one isn't: it's not a matter of degree, like wit or intelligence. How, then, can what is not a matter of degree at all be invoked to account for a disposition to accept an equal division in what necessarily are matters of degree? For the equal divisions claimed to be generated by MRC are genuine cases of equality: that is, we have divisible magnitudes of which each gets an equal share.[3]

That it is unreasonable to expect any other rational person to be willing to make a concession if one would not be willing to make a *similar*

[3] The distinction, which Gauthier points out is necessary, between cases of equal division and cases of unequal division which dominate the equal division in question is irrelevant here. See his discussion of Roth (p. 140, top).

concession is interesting, but on reflection invites the question, "Why?" After all, in bargaining contexts, we are arguing over how to cut up a piece of pie. More for you is less for me; but all we are given is that each wants *as much as possible*. If we say, "Yes, but how much *is* possible under the circumstances?" then the answer would seem in principle to vary depending on the propensities of the other bargainers. Given – as seems obvious – that anything *is* possible, within the range of just long of nothing (leaving almost all for the others) to just short of everything (leaving almost nothing for the others) for bargainer B_1, then the question of why we should be settling in the middle rather than somewhere else, as a supposed matter of "reason," becomes acute. The appeal to "equal reason" seems inappropriate.

What are the alternatives? One thought is this: if our bargainers are both recognizably human, then the fact that there is no inherent reason, in the nature of the case, why one division should be any better than another may be appealed to on the ground that disputes which cannot be settled by reason are likely to be settled by force; but the application of force in human affairs violates the Hobbesian equality principle. Hobbes, as we know, argues for equality in the "natural condition of man." He qualifies this almost immediately, noting that "there bee found one man sometimes manifestly stronger in body, or of quicker mind than another . . . "; but his conclusion is equalitarian nevertheless: ". . . when all is reckoned together, the difference between man, and man, is not so considerable as that one man can thereupon claim to himselfe any benefit, to which another may not pretend, as well as he."[4] The fact that you would win in a fair fight today may give you a temporary edge; but what about the unfair fight that will follow that one? And what about the fight in which my friends back me up? (And/or yours you?) And what about the question whether I should ever deal with you again, if I can possibly help it?

A further thought: we bargain, in the real world, against a time budget. Neither of us can wait forever, both can wait quite awhile, and all of the time spent waiting is time wasted, in principle. Assuming that our patience and, in general, our budgetary constraints in terms of time to spend bargaining are pretty equal, we would again get equality as a plausible operating assumption. The equality of this budgetary constraint is by no means axiomatic, though, and it is noteworthy that sometimes a bargain will be concluded unequally due to the superior patience of one party. Is Gauthier to claim that in these cases there has been injustice? But my man in the Arab market *owns* the carpet in question: he doesn't have to sell on *any* terms if he doesn't want to. And I

[4] Hobbes, *Leviathan* (New York: Dutton, 1950), p. 101.

own the money, which I don't have to part with if I don't want to. There is no right answer to the question: What is the *correct* price of the carpet?[5]

Moreover, we should recall that ordinarily the cooperation in question involves effort; and where there is a cooperative surplus, it is plausible to invoke the principle that one's share of the product should be proportional to one's share of the production cost, such as effort. To be sure, this intuitively plausible suggestion, which may be taken to be the root idea underlying the Labor Theory of Value of Marxist fame, runs into horrendous conceptual problems which can be resolved only by leaving the determination of what constitutes "equal effort," and the like, to the market.[6] Even so, it will often be possible to make a rough comparison of our respective efforts in relation to production, and then one can appeal to equal contributions as a basis for equal shares.

Perhaps, then, the principle of MRC is slightly misstated, or at any rate, Gauthier's account of it is misleading. In his account, it is not the disposition of minimax relative concession *itself* that makes for an equal outcome, but rather the assumption of "equal claims for equal rationality." It is *not* true, just like that, in the case where "a single transferable good, produced in fixed quantity and divisible in any way among the cooperators" is one such that "maximum concession is minimized if and only if each person receives an equal share of the good" (p. 153). What is true is only that my (or anyone's) maximum concession is minimized *relatively to the others – and that only on the assumption of a linear utility* function for the rational agent in question against quantities of the good in question – if and only if it is equally divided. "MERC" would be a better name for it.

However, let us put aside these quibbles, especially since the principle Gauthier proposes is so beautiful and so intuitively right. Instead let us address a very different and extremely important issue: What, in fact, are its implications? Now, some have evidently thought them considerable. David Braybrooke, for example, supposes that MRC would apply to a whole society, in which case we would have to try to figure out what the maximal and minimal claims of each individual would be – "making fantastic demands on information, which no contracting parties and no current critics of government could ever meet."[7] But, as I have complained elsewhere,[8] *it is by no means evident that MRC applies at*

[5] The relevant formal principle here is, I suppose, Zeuthen's, discussed by Gauthier on pp. 71–5. This principle says that the person whose ratio between cost of concession and cost of deadlock is less must rationally concede to the other. The empirical surmise is that considerations of human time budgets make it generally implausible to suppose that differences in this respect will make concession rational.

[6] Among many recent destructive discussions of the Marxian notions of value, I would mention my own in "Marxism: Hollow at the Core," *Free Inquiry* (Spring 1983): 31–2.

[7] David Braybrooke, "Social Contract Theory's Fanciest Flight," *Ethics* (July 1987): 760.

[8] Jan Narveson, *The Libertarian Idea* (Philadelphia: Temple University Press, 1989), p. 195.

this level, despite Gauthier's suggestion that "Society may be viewed as a single cooperative enterprise" (p. 274). For one thing, as Gauthier himself observes, when there are more than two persons, each receives "only that part of the surplus to the production of which he would contribute" (p. 134). And it is simply not plausible to suppose that everyone contributes to the production of everything in our or any society. Or if we say so, then the "contribution" in almost every case will consist merely of that person's refraining from upsetting the productive applecarts of those actually engaged in cooperative productive activities. And to regard *that* as a "productive contribution" is to bring us to our other subject, pursued in the following.

Meanwhile, there is one further point, possibly of fundamental importance. Gauthier contrasts two interesting cases: (1) that of Ms. Macquarrie, the pharmaceutical chemist, versus her lab assistant, Mr. O'Rourke, and (2) that of Sam McGee, the Yukon prospector, versus Grasp, the banker. Ms. M. discovers a wonder drug that makes her a millionaire, but does not divide her royalties equally with O'Rourke. Sam, on the other hand, forced to borrow a measly $100 from Grasp in order to register his claim to the richest vein of gold in the Yukon, "will (rationally) have to offer Grasp a half-share in the claim" (p. 153). The difference seems striking.

Why the disparity? Because Ms. M., it seems, did not carry on her experiments *as* a cooperative venture with O'Rourke. "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" (p. 153). Sam, on the other hand, enters into full cooperation with Grasp, without whom his venture will not succeed at all. And one thinks here of resourceful marketers who reap the major share of profits on inventions technically far beyond their capabilities, gleaned from cooperative ventures with the computer innovators, chemists, and so on whose ingenuity provides them with the goods they sell.

If the difference in the two cases seems fundamental, we must also admit that it is fortuitous. In another possible world, O'Rourke would be the only lab assistant available and without him Ms. Macquarrie could do nothing, and Sam could choose among a half-dozen sources for the loan he needs, as Gauthier agrees (pp. 153–4). Now consider ordinary business deals or the activities among any cooperating set of business people. How do they proceed? In general, by someone having an idea, getting in touch with others, making an offer, and so on. People to help are either hired on a wage/salary basis or taken on as partners at some level or other, or both; perhaps market shares are sold, and so on. People have a sense of when they have made a good deal and when not; and sometimes a deal is made in bad faith, as a result of deception, or under

some subtle or not so subtle type of coercion, in which case its effect is rejected or disputed. But when a deal is not defective in such ways, where does MRC, or anything like it, come into it?

Gauthier's characterization of MRC as a principle of "justice" puts it on the same level as, say, the principles of promise keeping, truth telling, or fair dealing. In fact, he suggests that all of them "are to be defended by showing that adherence to them permits persons to cooperate 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" (p. 156). Can we accept this? I think not. The validity of a promise is surely not due to its approximating an equal division of anything. If the background conditions are properly observed, and everything is on the up-and-up, then promises and contracts are valid by virtue of their form, of the fact of agreement itself, and not in virtue of the resulting distributions exemplifying some or other proportions. That does come into the parties' estimation of how satisfactory a deal they have struck: if it doesn't look promising, they don't make the deal. But it does *not* affect their sense of whether it was a valid agreement at all. If I later discover that Mohammed would have settled for $20 less had I but known, or if I decide to settle – gullible tourist that I am – on his opening price, then he doesn't owe me a dime. Should the product be not as advertised, or subtly defective, that's another matter. He is then guilty of deceit. But deceit isn't a matter of failing to approximate MRC in any sense I can readily think of.

It must also be evident that MRC makes sense only against the background of independent rights of the parties concerned: rights to their own persons in the way of abilities and other resources, and rights to assorted items of external property. If we don't already have those rights, no intelligible sense of bargaining can get off the ground.[9] Yet if we do have them, then the validity of the transactions we make concerning them is due simply to the continued operation of those rights themselves. You can give me an X because it is yours; I can give you a Y in exchange because it's mine. What makes Y yours when I do give it to you is simply the fact that I have done so, and not any supposed fact that the resulting distribution of some set of benefits

[9] See the important passage on p. 222: ". . . the emergence of either cooperative 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." Obviously, however, if this viewpoint is pressed too hard, then there is an end to the contractarian project as I have described it at the outset of this essay. See my brief discussion in Narveson, *The Libertarian Idea*, p. 190.

would approximate an equal proportion of our maximal preagreement claims.

In fact, given broadly Lockean rights in ourselves and in external items of property, it would seem that MRC is a subordinate and largely dispensable principle, best not regarded as a principle of *justice*, strictly speaking, at all. The moral force of actual bargains is transferred to the results of those bargains directly from the antecedent rights of the parties concerned to the goods or services being exchanged, rather than from their "claims" via Gauthier's interpretation of MRC.

And for good reason. For the principle we *really* need, once property rights are in place, is not MRC, but constrained maximization (CM). CM ensures that we both benefit from exchanges, by ensuring that we both actually get what we have agreed upon. It does not go into the question how *much* we benefit, leaving deliberations on that matter to the antecedently operating practical reasoning of each. MRC serves as a plausible guide here, to be sure; but that seems to be all.

Gauthier on predatory gains and the Lockean Proviso

Gauthier's version of the Lockean Proviso on acquisition forbids the pursuit of advantage by imposing disadvantage on others. The proposed baseline for assessing my advantage and disadvantage in interacting with person *P* is how things would be for me in *P*'s absence. If my action is such that I would do as well or better if *P* weren't around at all, then I am not taking unfair advantage of his presence. If, on the other hand, my action will lead to my benefit only if *P is* around, yet renders *P* worse off than if I had not been around, then I *am* taking unfair advantage of *P*.

The present question concerns what we should take as the baseline for our negotiations regarding future arrangements of society. Would we accept a status quo incorporating differential predatory gains of some parties due to past interaction? Or should we, as Gauthier insists, "purify" our starting point for this fundamental purpose by invoking the Lockean Proviso retroactively?

Gauthier had previously adverted to the matter with this thought: "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-cooperative interaction, if she is to find it rational to cooperate. 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" (p. 133). The question to be pursued here is: Why not?

For analytical purposes, at least, let us make a distinction between two sorts of "powers" that might conceivably be relevant to distributive

questions: what we will call "productive" and "predatory" powers.[10] Of course acts of predation are all, in their way, productive: if *A* makes his living by stealing from *B*, then he is, if successful, producing something, namely, what he takes to be a better overall situation for himself, or whoever is the recipient of the ill-gotten gains in question. However, those gains had originally to be "produced" in a more fundamental sense of that term. When *A* gets on by predation on *B*, what *A* does is to redistribute wealth, whereas *B*, let us assume, was the actual originator of it.

It is obvious, I take it, that contributions to production, in this latter sense, *are* a relevant basis for distribution. Possession of unusually great productive powers, then, is a likely basis for some sort of distributive recognition.[11] It is less obvious that predatory powers are so. One might simply insist that they are no basis whatever for just distributive claims, and indeed, both Hobbes and Locke would agree with this in one way or another.

Hobbes, familiarly, imagines a prepolitical and premoral "State of Nature" in which distributions of goods are determined by the total array of powers, predatory as well as productive, available to each party. But Hobbes claims that the distribution of predatory powers is such as to render the overall distribution in the State of Nature equal, as we have already noted before: "... when all is reckoned together, the difference between man, and man, is not so considerable as that one man can thereupon claim to himselfe any benefit, to which another may not pretend, as well as he."[12] In Hobbes' view, the plausible outcome is to forbid predation across the board, leaving the question of how to distribute the produce from productive powers to the Sovereign to decide.[13]

But if it is plausible to accept Hobbes on the rough equality of predatory powers, the distribution of creative or productive powers is another matter. These, surely, are very unequal. In Locke's view, in contrast to Hobbes', the proper principle regarding the distribution of goods flowing from the exercise of these productive powers is, in effect, to leave them to the market, constrained by property rights.[14] This is the Libertarian's view, and the one to which I am inclined.[15]

[10] This should not be confused with C. B. Macpherson's distinction between "extractive" and "developmental" powers, which is, I believe, essentially confused. See *Democratic Theory* (Oxford: Clarendon Press, 1973), pp. 40–52.

[11] There are social philosophers who apparently deny this, and Rawls may be among them. I will not discuss their arguments in this treatment.

[12] Hobbes, *Leviathan*, ch. XIII, p. 101.

[13] "Seventhly, is annexed to the Soveraigntie, the whole power of prescribing the Rules, whereby every man may know, what Goods he may enjoy ... without being molested by any of his fellow Subjects: And this is it men call *Propriety*." Hobbes, *Leviathan*, chap. XVII, p. 149.

[14] Locke, *Second Treatise of Civil Government* (New York: Dutton, 1966), chap. v, pp. 129–

But a question arises about the relevant principles here. Suppose that predatory powers are *not* equal: then what? Gauthier introduces his deliberations on this matter with a tale about a society of masters and slaves, the latter serving the former through coercion (p. 190 ff.). Only force, rather than any sense of moral obligation or ideology, preserves the situation. However, we are to suppose that the masters currently have the whip hand: the coercive apparatus by which the masters remain masters is essentially permanent, the slaves being incapable of unilaterally dismantling it. Still, the costs of maintaining the apparatus are considerable to the masters. They therefore propose a deal: "You slaves will continue to serve, but *voluntarily;* in return, we masters will refrain from beating you, and even give you better living conditions into the bargain. What do you say?" According to Gauthier, this offer is one that cannot be rationally accepted, because its outcome could not be rationally complied with. He puts the following speech into the mouth of the new Prime Minister, elected by the Ex-Slaves: "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 cooperation 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 . . . " (p. 191).

If this is the correct reasoning, then Gauthier has served up a lesson for all masters: since you can't expect stable optimal bargains from slaves, don't bother to try to treat them better in hopes of improving your own situation through cooperation! But is the PM's reasoning correct? Or is there, instead, hope for slaves, short of chancy resort to violent revolution? Let us consider.

Gauthier points out that "fair procedures yield an impartial outcome only from an impartial initial position" (p. 191). True. But he has also made it clear that *impartiality* does not imply *equality*. The division of profits between me and Robert Redford from my movie, in which he gets six million while I settle for two hundred thousand, is fair and impartial and sanctioned by MRC, though of course not equal.[16] The question, then, is whether the particular inequalities of the initial situation prior to the acceptance of the Social Contract, in which many of those who are better off are so by virtue of having exerted coercive force on those worse off, are capable of being incorporated properly in a baseline for fair bargaining.

40. To further complicate matters, the realization of benefits from the exercise of these powers is due largely to cooperation. Consequently, important issues arise concerning the distribution of those benefits. The discussion in the fifth section of this essay is concerned with that issue.

[15] See Narveson, *The Libertarian Idea,* esp. chaps. 6–8 and 15.

[16] Except, as we have seen, in the peculiar way that minimax relative concession reckons.

As Gauthier correctly observes, predatory activity is suboptimal. (We must distinguish between the situations under discussion here and, for example, those of sadistic masters who simply enjoy coercing their (non-masochistic) slaves; the latter will not be considered here.) The efforts of the coercers must be unproductive so far as they go: they stimulate productive activity on the part of the slaves, but, we assume, the slaves would be capable of equal or superior production in the absence of force, if they could be sufficiently motivated by other means (such as the prospect of a reasonable share in the product). The time now spent in beating the slaves could be spent more pleasantly by both parties: playing chess, perhaps, or taking a nap. Why, then, shouldn't both parties accept, and keep, an agreement in which the same productive/service activities as before are carried out, but the time formerly devoted to the exertion of force by the masters over the slaves is now devoted to these more desirable activities? It would be better for both parties, we are assuming. So why not?

It is important to bear in mind that in our contractarian deliberations, we follow Hobbes in assuming that the State of Nature is a *totally amoral* state – a condition in which *nothing* is wrong, nothing forbidden. Contrary to the assumptions of our story, Hobbes also believed that this initial state was one of equality of coercive powers. We are in part here exploring the implications of the non-Hobbesian assumption. If we make the Hobbesian assumption of equality of coercive powers in the State of Nature, then it is plausible enough, I take it, that we can get from Hobbes to Locke: the use of force to pursue one's ends, in all cases when a cooperative arrangement could have yielded a greater benefit, will be unequivocally and universally forbidden. But it is less obvious, at least, that this will be so if our starting point is one of *unequal* predatory powers.

Many may find it not so clear that we can get, as I put it, from Hobbes to Locke. However, I am here exploring the situation if we are given the Hobbesian assumption of roughly equal predatory powers in the status quo ante. Still, it might be argued that the flat prohibition on force might well be less plausible than various alternative formats in which *some* ends could be pursued using coercion, at least if it were employed by a public agency under fair rules. This is not the place to pursue this important subject. For the present, let us assume that there is a flat prohibition on privately wielded force for private ends other than self-protection, leaving the more difficult question for another time.[17]

Specifically, let us consider the case where *A* is greatly superior to *B* in predatory capabilities, while *B* is greatly superior in productive powers. *A* is accustomed to making his living by raiding *B*, who understand-

[17] I explore this matter, if not as satisfactorily as I would have liked, in Norveson, *The Libertarian Idea,* chap. 14, pp. 154–84.

ably finds this a nuisance. Against this background, *A* proposes a deal: *B* will simply transfer to *A* a fixed percentage of the fruits of *B*'s highly effective labors, and *A* will thenceforth renounce the use of force in his dealings with *B*.

There are two cases to distinguish. In case 1, *A* puts his weapons in a convenient place, to be taken out instantly in case B should fail to live up to his part of the bargain. In case 2, however, *A* melts his swords into ploughshares, which he cheerfully supplies to *B*, whose productivity is thereby increased the more. Since *A* gets a percentage of *B*'s product, he naturally expects to share in the benefits.

If Gauthier is right, case 2 is one in which *A* is in for a nasty shock. B, if rational, will cease to deliver as soon as the last sword is duly converted into the last ploughshare. Obviously there is a complication in the example in that the *technology* of warfare presumably has not been melted down along with the hardware. In a real-life situation, there would be the likelihood that *A* would be able to resume full-scale predation in fairly short order. I therefore wish to rule this out for present purposes, admittedly at the expense of making the example still more artificial than it already is: my swords-to-ploughshares conversion is irreversible. My question now concerns the justice of *B*'s Gauthier-sanctioned action, given Gauthier-sanctioned premises (which I in general share).

The situation is made yet more interesting if we suppose that negotiations are carried on between *A* and *B* in which the distinction between cases 1 and 2 is very clearly and explicitly recognized by the parties. Importantly, we want the move from 1 to 2 to constitute, in the abstract, an optimal strategy: both parties are better off in 2 than in 1. But at the time of the negotiation, *A* is in a position to choose between them, and *B* is aware of this. Why wouldn't this agreement be binding?

As I understand it, Gauthier's reason for thinking that it would not be rational for *B* to carry out this agreement is that at the time at which it is to be carried out, the coercive apparatus by means of which *A* was able to exact agreement in the first place has been dismantled. In his fable, the newly elected Prime Minister, an ex-slave, says, "Once that power was taken away, it became obvious that the fruits of cooperation 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." To which Gauthier adds, "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 . . . " (p. 191).

But if party *A* (the masters, in Gauthier's example) foresees that party *B* (the slaves) would use this reasoning, then of course they will not make the deal. For we are supposing that they would be worse off under noncompliance by the (ex-)slaves than under the status quo, and Gauthier agrees that where that is so, the rational masters will not move voluntarily to the new situation.

Gauthier's example is clouded by the Prime Minister's further remark that the Masters had probably "saved themselves a revolution" in trying their deal, despite its outcome having been very contrary to expectations (p. 191). But let us assume that a revolution mounted against the Masters with their current coercive technology intact would be futile, whereas a revolution mounted after the agreement would, indeed, be successful – but only because the Masters had *voluntarily dismantled* the technology in question.

Now, one would think that they *would* move voluntarily to a situation which, after all, dominates the status quo – would, that is, if this new situation is possible. Its possibility, however, is contingent on compliance. And if the argument is that compliance is impossible, then there is bad news awaiting us. For consider the principle of constrained maximization itself, which calls upon us to forgo noncooperative gains in Prisoner's Dilemma situations. If we look at agreements generally, this attitude would seem to land us in Hobbes' "easie truth," that "Covenants are but words, and breath, having no force to oblige, contain, constrain, or protect any man, but what it has from the publique Sword."[18] The person who carries out her part of a bargain first, lays herself open to noncompliance from the second party, who has now got what he wants. Why shouldn't he take the money and run? Yet that is precisely the attitude that constrained maximization rules out. Why should things be any different when the Pareto superior situation is a move up from an arrangement originally secured by superior force alone? Why, that is, when the situation ex ante was one in which the agreement in which "Morals by Agreement" consists is *not* as yet in place. Doesn't that make it out of order to object to the initial depredations as "unfair"?

Gauthier does not, as he cannot, in general disavow the use of predatory force in interpersonal relations. Consider the rather significant subject of our dealing with animals, for instance. We do not make a Social Contract with those individuals. One reason, no doubt, is that we can't: animals are not in general capable of entering into contracts

[18] Hobbes, *Leviathan*, chap. XVIII, pp. 146–7.

at any very high level of sophistication, we suppose. But that isn't the only thing blocking the way to a General Mammalian Contract. The fact is that we (or at least, we carnivores) regard our present situations as preferable to those we might be in were we to constrain our behavior toward them by invoking the Lockean Proviso. The benefits are not mutual. Were the Lockean Priviso a categorical imperative, we would have to resort to the implausible move of the likes of Kant and Descartes, holding that non-human animals are "irrational," in order to justify our rather predatory behavior toward them. But we don't need to do any such thing, nor does Gauthier. We simply say, "Tough!" And we do not think we are thereby violating any duties we have toward them. In the view of us contractarians, there simply are no such duties. The foundation of duty is mutual advantage, and in the case of the animals, a contract, were one possible given current animal capabilities, would not be advantageous to us. But why is not mutual advantage a sufficient, rather than merely a necessary, condition of the rational assumption of duties? And is not the relation of slaves to masters precisely one where mutual advantage is an obvious possibility? Consider the workhorse, broken to the harness in manifest opposition to its native instincts. Still, the horse may reason that it is better off if it simply obeys orders and collects its daily allowance of oats and water than by kicking against the pricks and retaining the disposition to run off whenever its master isn't around.[19]

Second: we identify predation with the use of force. However, I wonder whether we might not also consider the case of superior powers generally as instructively akin to it, though of course not identical. Suppose that I am extremely fond of piano music, and that a certain individual, Vladimir, can supply my demand far better than I can myself (I'm a hopeless pianist). Suppose too that Vladimir is a born pianist, and I am not. We have labored equally, or perhaps he far less than I, but, owing to inborn ability, Vladimir has become a great pianist, whereas I am somewhere around the write-off level. There is nevertheless no objection whatever to his entering into exchanges with me that are, at least in some points of view, greatly to his advantage. Like Wilt Chamberlain, Vladimir cheerfully collects a modest fee from thousands of admiring music-lovers, thus accumulating wealth beyond the wilder dreams of mere professors. Now along come certain social philosophers who object to this accumulation based, for present purposes, on the

[19] As Gauthier points out in connection with workhorses, for instance, on p. 17. For further defense of this point of view, see Jan Narveson, "Animal Rights Revisited," in *Ethics and Animals*, edited by Harlan Miller and William Williams (Clifton, NJ: Humana Press, 1983), pp. 19–45. An important commentary, in a considerably different vein, is in the essay following mine in that collection, Annette Baier's "Knowing Our Place in the Animal World," pp. 61–78.

accidents of fortune. Why, they say, isn't Vladimir unfairly exploiting my lack of powers (and, of course, my inordinate passion for his wares)? Why shouldn't a Lockean Proviso forbid inequalities of all sorts that are due to the mere accidental distribution of powers, be they predatory or benign?

The reply is simple and convincing. Vladimir's "exploitation" of me, if that is what anyone wishes to term it, is to my advantage, whereas predatory exploitation – that is to say, exploitation properly speaking – normally is not. In exerting his superior powers, Vladimir contributes to my overall level of satisfaction, whereas the predator, in exerting his, detracts from it. It is a good answer, and the right answer. But there is a problem. In the hypothetical Hobbesian State of Nature, neither of us cares anything about the other, and if you, due to superior predatory powers, are predatorily exploiting me, that's just too bad for me. But then, that may also be the background for a bargain: you cease your predatory violence and I make it worth your while. If we then develop cooperative relations still more to our mutual advantage, as well we may, then bully for both of us! Do these incorporate in some unfair way your initial predatory gains? Must the Social Contract begin by requiring compensation from you to me for all those gains exacted by superior force? I am inclined to be dubious about this.[20]

Perhaps some suppose that these gains hang around, infecting all future transactions in an unsatisfactory manner. I think this unlikely, but in any case it doesn't clearly prove anything. Short-term consumption gains would, of course, not "hang around." So what about capital gains, so to speak? They might indeed continue into the indefinite future. But it is unclear whether this matters. For one thing, once we have respect for property rights in place, then if I am ingenious, my gain could be greater than yours in any case, for I may convert my initially lesser resources into a capital greater than yours in the process. More importantly, however, there is the fact that as time goes by, both of us are better off. Not only are we better off than we would have been had we remained in the State of Nature, but if all goes well, each of us is better off at any given time than we were in the preceding time. Even if you, who had the superior starting point by virtue of your superior predation in the status quo ante, remain ahead of me, I remain ahead of where I would otherwise have been.

We must also remember that capital is only capital. In order to derive continuing rewards from it, there must be some whose labors are rendered more efficient by its use, and normally most of them are persons other than ourselves. Our profit from ownership, therefore, will derive

[20] As Gauthier has noted, James Buchanan, in *The Limits of Liberty* (Chicago: The University of Chicago Press, 1975), takes a contrary view to his. See esp. pp. 23–5.

from a *further* agreement with those others regarding their remuneration for this labor. The minimum claim of the laborers will be set by the condition they would have been in had we remained in a predatory condition. But their maximum claim is quite another matter. If our labor makes you, the erstwhile predator, much better off than you were in the predatory condition, then it is also rational for you to agree to a division of it enabling us to share in these further gains. Otherwise we would have no incentive to provide this greater gain; and yet, since it is above the level you would have been in had you remained in the State of Nature, it is not rational for you to threaten a return to it (as Gauthier rightly observes).

Once the Lockean Proviso sets in, our benefits from cooperative activity follows the rule To Each His Marginal Product, and in all probability, the effects over time of initial inequalities due to previous predatory gains become trivial. (Peter Berger observes that the upward social diffusion of the bourgeoisie in Europe following the industrial revolution was "probably affected above all by the intermarriage of aristocratic men and bourgeois women (the typical bargain by which the latter obtained a title for themselves and their children, while the former were rescued from bankruptcy by the dowry bestowed on daughters by doting bourgeois fathers)."[21]).

Thus, we should not accept the argument that in settling on a beginning point that incorporates predatory gains, we are necessarily setting a bad precedent for the future. That future will, indeed, contain no predation, because we will accept as part of the initial bargain the Lockean Proviso for the future. Our predator dismantles his predatory apparatus permanently, for he realizes that in inducing me to arm myself against his predation both he and I are wasting efforts which could instead be converted into mutual gains. The initial predatory gains, however, were exacted *prior* to the onset of our Social Contract. At the time they were exacted, to exact them was not wrong, because nothing was wrong. To take the view that we must retroactively extend the Lockean Proviso to cover all past history is to suppose that morality is natural in a stronger sense of "natural" than Gauthier can accept. For he is, after all, advancing the thesis that morals are "by agreement."

On the Hobbesian starting point

When the Social Contract view is advanced as a set of propositions in abstract decision theory, as it is by Gauthier, then there is a temptation to argue that the "State of Nature" is purely an abstraction, with no possible real-world instances to concern ourselves about. But this had

[21] Peter Berger, *The Capitalist Revolution* (New York: Basic Books, 1986), p. 99.

better not be true. The contract is not, indeed, historical in the sense of a universal meeting of all persons at some given time. But if it is to have any real-world significance, it must be something that can happen in particular minds at particular times. And it is possible that there could be encounters between individuals neither of whom had internalized the principle of constrained maximization. A sequence of such encounters could lead to a mutual appreciation of the advantages of peace, and thus to a real bargain between those erstwhile warring parties, followed by a real-world instantiation of the very problems I have just been discussing in the slightly fanciful terminology of classical social contract theory.

Hobbes may have thought that there was actually a time when people had no morality at all. But he also thought that morality presupposes government, and the absence of government is at least more obviously a possibility, one might suppose, than the absence of morality, the lack of operative social rules of any kind at all.[22] Nevertheless, let us try to imagine a circumstance in which people regard human predation upon fellow humans with roughly the same indifference that they so regard animals' predation upon fellow animals. When *A* kills *B*, depriving B of *B*'s hard-won goods, *C* merely looks on curiously, or in turn sets about attempting to get it away from *A*. In general, the situation would be that no one applies any sort of reinforcing epithets, nor engages generally in any other recognizable moral reinforcement. The baseline for morality in such a condition must, surely, be wherever the parties are at the time they began to appreciate the advantages of moral restraint.

All of this reflection has been on the assumption that what we have regarded as "gains" are not measured in terms of what Hobbes calls "Glory." If this happens, of course, we get into a zero-sum game (your greater glory is necessarily my lesser and vice versa), and then all bets – all hope of a mutually beneficial Social Contract – are off. In this respect, perhaps, we may accept the charge that all contractarian speculations are infected with "bourgeois values." If the overriding passion of our lives is not simply to get ahead of where we were before, but rather to get ahead *of the others*, no matter where those others or we ourselves may currently be, then gains from cooperation are impossible. That vast majority of us who (I suppose) share in the "bourgeois" orientation which settles for a good life, reckoned in terms of values that require no such comparisons, have a great interest in pointing out to any who

[22] R. E. Ewin, in his important though little noted book *Cooperation and Human Values* (New York: St. Martin's Press, 1981), points to the overwhelmingly fundamental role that certains kinds of cooperation play in human life, observing that it must be questionable whether there could even be language without cooperation, since discourse depends on mutual understandings that are examples of cooperation. See pp. 10–13.

suppose themselves to be otherwise in this respect that their predilections for glory will in all likelihood lead to familiar bad ends.

In the State of Nature, those who invade for gain are engaging in activities that are as "legitimate" as any others one might attempt. But in the civilized state in which we realize the gains of cooperation, invasion for gain is not legitimate, and when it occurs, it is strictly *unfair*. Those who live by plundering in the civil state have taken advantage of their fellows, who have extended, throughout the lives of the predators, their protection and forbearance – which would *not* be true in a genuinely Hobbesian State of Nature. Had they any reason to suppose that certain persons would take to plunder, they could have quickly rendered nugatory any threat of force from those quarters. The predator under the social contract, therefore, has no leg to stand on. He cannot maintain that he is entitled to his predatory gains, because there would have been no possibility of such gains had there never been any contract, and a reasonable contract would forbid predatory activities from the start. There is, therefore, no reason to worry about setting a bad precedent were there any initial predatory gains to recognize. The future is different from the past, and bygones when things were very different should be let bygones.

Relations between states

I have pointed to one area in which this theoretical discussion is more than merely theoretical, namely, our relations to the lower animals. Possibly another, and much more potentially important if so, is that of mutual relations among nation-states. There we have many complications – far too many to permit a conclusive discussion here. Moreover, we should surely reject the claim that nation-states are simply in a State of Nature. All or nearly all states have made gains because of the willing cooperation of other states – at the very least the cooperation constituted by forbearance from exploitative activities, but usually a good deal more as well. In general, that is, states have *not* been in an antecedently amoral mutual condition. Nor do modern states claim to have been.

Besides, the very existence of states is surely suspect on the contractarian view. Once we see that the Sovereign is not the answer to the initial problems posed by the Hobbesian State of Nature, we may also see that it is not the answer to those posed by civilized society, either. If all individuals deal with all other individuals *as* individuals, then the supposition that aggressive war can be waged to advantage quickly reduces to an absurdity. That is an important conclusion, surely. Still, it remains that the history of the modern world has been characterized by the emergence of groupings of states founded on military alliance and mutual noninterference pacts, the object being "collective security"

– against other similar alliances or against individually formidable states. It is hardly to be doubted that these have often been characterized by inequalities due to prior predation. And global peace was thought to be contingent on a rough equality, especially military equality, between the major coalitions. The system was (is?), of course, unstable and not very effective. But what interests us here is that its subordinate members, while very keen indeed on obtaining political independence, did not always press for the reimbursement of gains made by the coalition leaders in the process of putting together their empires. Mexico has not insisted on the return of Texas; and so on. Sometimes, to be sure, they do so insist, and sometimes an uneasy or "cold" peace ensues when hostilities cease, but issues about land claims are not really settled.

The point here is that insofar as wars and other international disagreements are instances of unilateral abrogation of reasonable moral constraints, it is not clear that Gauthier's rule is going to make for peace. Certainly it will not in the short run; in the longer run, the question is more difficult. One can argue that if we insist always on the Lockean rule, the message conveyed will be that war is futile, since no long-run gains from it are possible. But the alternative message is that any nation that does make such gains is going to have to remain indefinitely on a war footing in order to hang onto them, since as soon as it disarms it will be forced to divest itself of those gains. Can there be any doubt at all that at least in material terms (and numbers of lives lost), the Middle East Arabs would be better off today had they simply made peace with Israel at the start, despite the latter's arguably wrongful incursions into their territory? Material benefits aren't everything, of course. But it should surely be matter for doubt whether the particular kind of "nonmaterial" values that reinforce bloody and intractable conflicts are all that terrific.

But that gets us into areas lying beyond the confines of this investigation. As said before, the discussion here does not pretend to be conclusive, but only to cast some doubt on an interesting and important feature of Gauthier's work.

Summary

The persuasiveness of the contractarian theory lies in the fact that it secures the weight of our nonmoral reasons on behalf of moral constraints. Only thus can it claim that *any* rational person will be bound by these considerations. In order for this crucial aspect of the theory to be effective, we must show that the envisaged "starting points" for the theory are ones about which we must be concerned. We must be sure that the contractarian machinery incorporates no spurious normative inputs in the process of generating moral results. This concern motivates

queries about Gauthier's arguments for minimax relative concession, on the ground that the basis appealed to – "equal rationality" – appears to be in principle spurious, though some nonspurious (but uncertain) reasons are suggested. I have also queried Gauthier's condition for the retroactive application of his fundamental principle against the use of force and fraud, the Lockean Proviso. Though fully accepting the proviso itself, doubts arise concerning its application, which denies continued possession of their gains to those who previously gained from predation. This would seem to preclude benefits to some parties under some conceivable conditions. I then asked how this argument would apply to real-world situations of war and peace. My argument here is that extending civilized protections to all depends on taking Hobbes seriously, in particular on the point that we are all essentially equal in our capacity for predation. That assumption, plausible at the level of one individual against another, is not obviously so at the group level. I do not claim to have resolved this issue, which may appreciably effect our understanding of the contractarian theory.

10. Equalizing concessions in the pursuit of justice: A discussion of Gauthier's bargaining solution*

Jean Hampton

In *Morals by Agreement*, David Gauthier embraces the Hobbesian thought that morality is authoritative for us only insofar as it advances our interests. But this thesis on authority has implications for the content of morality: if morality is to further our interests, then, says Gauthier, it must be possible to generate it "as a rational constraint from the nonmoral premises of rational choice." But how can this be done? Gauthier's problem is to explain, first, how unattached, mutually unconcerned utility-maximizing individuals whose interests frequently conflict can come to agree on the *terms* upon which it would be rational for each of them to cooperate with one another; second, how they could be trusted to comply with these terms; and third, how to define the initial position from which cooperation should proceed in order for the resulting distribution of cooperative benefits to be fair. His solutions to all three problems are controversial, but here I want to evaluate his attempt to solve the first problem by defining terms of cooperation using what he calls the principle of minimax relative concession (hereafter called the MRC principle).[1]

Gauthier elaborates on and defends this principle in Chapter V of *Morals by Agreement*. Consider, he says, that in the first stage of a bargaining process, each party advances a claim. If (as is likely) these claims are incompatible, there is a second stage in which each party offers concessions to the others by withdrawing some portion of his original

*Excerpted with minor additions by the author by permission of the *Canadian Journal of Philosophy* from "Can We Agree on Morals?" by Jean Hampton, in *Canadian Journal of Philosophy* 18 (1988): 331–56.

[1] I say something about his solution to the second and third problems in "Two Faces of Contractarian Thought," Chapter 3 in this volume, and focus on his solution to the second problem in "Constrained Maximization and the Nature of Reason," unpublished manuscript.

nomic environments as objects of distributive mechanisms. An economic environment is defined by an n-dimensional space of commodity bundles to be divided between two persons and the utility functions of the persons on the space of bundles of these commodities. He investigates how classical bargaining axioms work when they are reformulated as axioms on economic environments. Moreover, Roemer[34] presents some axiomatic characterizations of distribution mechanisms using economic information. Roemer's contribution clarifies under which conditions the investigated mechanisms correspond to some of the classical bargaining solutions. One should mention that in his models, Roemer is assuming interpersonally comparable utilities, whereas the models discussed in this essay do without any degree of interpersonal comparability.

Roemer draws the conclusion that only if the dimension of the commodity space is unbounded, the classical bargaining solutions (with interpersonally comparable utilities) can uniquely be characterized by axioms using economic environments, and he does not consider this condition as appropriate for economic bargaining problems.

[34] J. E. Roemer, "Axiomatic Bargaining Theory on Economic Environments," *Journal of Economic Theory* 45 (1988): 1–31.

Part III

The rationality of keeping agreements

Overview of the essays

In his essay Geoffrey Sayre-McCord argues that Gauthier fails to establish that rationality always, or even almost always, requires human agents to dispose themselves to comply with the requirements of morality. For in the real world, people's dispositions are opaque enough that it is often possible to deceive others into thinking that one is trustworthy. Given this possibility of deception, rationality dictates, at least for some real-world agents, that they adopt the policy of pretending to be trustworthy, but breaking agreements whenever it is to their advantage.

David Copp reaches this same conclusion. He also questions Gauthier's unconventional view that a choice is rational if and only if it conforms to a disposition, the having of which would give the agent at least as much utility as any alternative disposition. The more usual view is that a choice is rational if and only if *it* would give at least as much utility as any of *its* alternatives. Copp acknowledges that even if the claimed connection between rational dispositions and rational choice is rejected, Gauthier's argument that it is rational to choose to be disposed to comply with rational agreements is still important. For, if successful, it shows that rationality requires people to become disposed to keep their agreements, and that is still a fairly strong result. Copp challenges, however, Gauthier's claim to have thereby established that rationality dictates that one comply with *morality*. Gauthier's argument for this claim rests on his equation of morality with rational and impartial constraints on the pursuit of self-interest. Copp argues that this equation is mistaken – at least for the sense of impartiality that Gauthier invokes. Copp concludes that Gauthier has not established that rationality dictates that we

be disposed to be *moral*. He argues further that if even Gauthier's argument were successful, it would not be an adequate answer to the moral skeptic. At best, Copp claims, it would show that behaving morally is sometimes rationally justified – not that morality itself is.

Holly Smith also criticizes Gauthier's argument for the rationality of keeping agreements. She argues that the advantages of constrained maximization over straightforward maximization are not as clear as Gauthier claims. Gauthier's argument rests on an illegitimate restriction of the policies open to both the agent and his/her potential partners. Moreover, even if all agents are constrained maximizers, and they are transparently so, it does not follow that they will always cooperate. Smith further challenges Gauthier's claim that if it is rational to adopt a given policy (e.g., that of keeping rational agreements), then on each occasion, it is rational to choose on the basis of that policy (e.g., keeping a given agreement). The rationality of adopting a policy, Smith argues, does not guarantee the rationality of a choice that conforms to the policy. Finally, Smith argues that in any case, Gauthier's argument does not succeed in deriving morality from a morally neutral foundation. Close scrutiny reveals that either his derived principles are not genuine moral principles or that his principles of rationality are not morally neutral.

Jody Kraus and Jules Coleman criticize Gauthier's argument that it is uniquely rational to be *narrowly* compliant with agreements (i.e., to be disposed to comply with the terms of *fair* bargains – but not necessarily disposed to comply with the terms of *unfair* bargains). Their main point is that under appropriate circumstances, it can be rational to be *broadly* compliant (i.e., disposed to comply with any mutually advantageous bargain – fair or not). Given that morality (at least on Gauthier's view) is both rational and fair (impartial), they conclude that Gauthier has failed to show that rationality requires that we be moral.

Like these authors, Peter Danielson holds that Gauthier's argument for the rationality of constrained maximization fails. Unlike these authors, however, he holds that it is only the details of Gauthier's argument – not the basic approach – that fail. He argues that there is a policy the adoption of which yields more utility than straightforward maximization, Gauthier's constrained maximization, or any other policy. This is the policy of reciprocal cooperation, which – simplifying somewhat – directs agents to cooperate with others when and only when this cooperation is necessary and sufficient for the cooperation of those others. Unlike Gauthier's constrained maximization policy, reciprocal cooperation directs the agent *not* to cooperate when interacting with others who will cooperate unconditionally (even if they know that the other agents will not cooperate). Thus, reciprocal cooperation yields all the benefits of cooperation that constrained maximization yields, and it yields some benefits from exploiting unconditional cooperators that the

policy of constrained maximization does not yield. Where there are just two agents, reciprocal cooperation, not constrained maximization, Danielson argues, is the rational policy to adopt. In the many-person case, however, things are more complicated, since the rationality of cooperation will depend on *how many* agents are willing to cooperate. Here Danielson argues that a policy called "counteradaptive cooperation" is appropriately sensitive to this feature, and is therefore the rational policy to adopt in the many-agent situation. Like Gauthier, then, Danielson holds that it is rational to adopt a policy of cooperation under certain conditions. He disagrees with Gauthier, however, concerning the nature of the conditions under which cooperation is rational.

A caveat: Gauthier holds that under a broad range of circumstances, rationality requires that we conform to the terms of agreements it would be rational to make. He also holds that an action is morally permissible just in case it conforms to rules that it would be rational for the members of society to which to agree. Consequently, he holds that under a broad range of cases, rationality requires that we behave morally. Here there are two distinct issues: (1) Does rationality require us to keep our agreements? (2) Does rationality require us to be moral? Given Gauthier's contractarian theory, the two issues are coextensive, but conceptually they are distinct. Since authors do not always clearly distinguish these two issues, readers should be careful to determine exactly which claim is being assessed.

egies as interesting as UCP and provides a way to test them, even though they may prove superior to reciprocal cooperation.

Conclusion

David Gauthier's concept of constrained maximization is a path-breaking attempt to solve the compliance problem, "the profoundest problem in ethics." Constrained maximizers are discriminating; they cooperate with and only with the similarly disposed. But it is not clear how similar a disposition it is rational and moral to demand. Other agents may be more, less, or exactly as cooperative as oneself. In each case, we have seen that Gauthier's preferred disposition is problematic. Constrained maximizers' agents cooperate with niave unconditional cooperators. I argued that this is not rational and that allowing others to exploit these innocent agents is also not moral. The third section showed that the problem of coordinating with agents very similar to oneself requires more procedural specification than Gauthier provides.

I conclude that reciprocal cooperation is more rational and no less moral than constrained maximization. Thus, reciprocal cooperation tentatively closes the compliance dilemma in the two-player Prisoner's Dilemma. This is a crucial testing ground for rational morality, but it is not the only test. Indeed, in retrospect, the Prisoner's Dilemma appears ideally suited for the combination of openness and opportunistic adaptation that enable reciprocal cooperators to prevail. Other situations, like Chicken and many-player games, may prove resistant to these techniques. For this reason, I have tried not only to specify a decision procedure that solves the compliance problem in the PD, but also to develop methods to generate and test new candidates for rational morality in other situations. We have seen that it is often hard to imagine new alternatives and difficult to test them due to strategic complexities. Both problems have their analogues in the field of artificial intelligence. I conjecture that artificial morality will prove to be a fruitful way to develop and test the seminal ideas of *Morals by Agreement*.

17 Rational constraint: Some last words

David Gauthier

The view ex post differs from the view ex ante. Critical response and further reflection would yield a different book. Despite the best efforts of some of my critics to convince me otherwise, I still think that most, and the most important, of the particular arguments that appear in *Morals by Agreement* (henceforth *MbA*) are sound. But when I wrote *MbA* I did not fully grasp the structure of the theory of which the particular arguments are the parts. I shall therefore take the opportunity afforded to me here of having (for the moment) the last word to say something about this structure.

I treat morality as involving *constraint*. I introduce this idea in the opening pages of *MbA*, claiming "to defend the traditional conception of morality as a rational constraint on the pursuit of individual interest" (p. 2). This formulation is unfortunately misleading insofar as it suggests that morality merely constrains egoism; I want to defend morality as a rational constraint on the pursuit of one's aims or objectives, whether or not these objectives have any connection with one's interest, or one's personal well-being.[1] Yet, of course, I also want to insist that one acts rationally in pursuing one's aims or objectives; the formal aim of the rational individual is the maximum realization of her substantive aims.

[1] Not only my reference to interest, but my insistence on nontuism, is misleading. Christopher Morris discusses some of the problems with nontuism in "The Relation between Self-Interest and Justice in Contractarian Ethics," in *The New Social Contract: Essays on Gauthier*, edited by E. F. Paul, F. D. Miller, Jr., and J. Paul (Oxford: Blackwell, 1988), pp. 154–72 (the contents of this volume appear also as *Social Philosophy and Policy* 5 [1988], with the same pagination); see my reply in "Morality, Rational Choice, and Semantic Representation: A Reply to My Critics," in the same volume, pp. 213–17. See also Christopher Morris, "Moral Standing and Rational Choice-Contractarianism," and Peter Vallentyne, "Contractarianism and the Assumption of Mutual Unconcern," Chapters 6 and 5, respectively, in this volume.

How then can a constraint on the pursuit of one's aims be rational? And if such a constraint is rational, is it therefore moral?

In *MbA*, I appeal to the structure of interaction made explicit in the Prisoner's Dilemma (henceforth PD) to show how constraint can be rational and how rational constraint can be moral. In a PD, each person has a strongly dominant action or strategy – that is, an action that is her best response (in terms of maximizing the realization of her substantive aims) to whatever be the action or actions of the other(s). Each does best, then, to choose her dominant action. But everyone would do better were each to choose some alternative action, and so to exercise constraint in pursuing the realization of her substantive aims. Each would do worse, of course, from her own constraint, but each would gain more from the constraint of the other or others than she would lose from her own.

Introducing the PD, and noting its ubiquity in interaction, naturally suggests that persons, expecting to face such situations, would do best, individually and collectively, to agree to mutual constraints. No one, seeking the maximum realization of her aims, would rationally undertake to constrain her direct pursuit of that realization unilaterally, but each should rationally agree to constrain herself provided her fellows do the same. We thus show how a constraint on the pursuit of one's aims can be rational. And since a constraint is rational only insofar as it is mutual, so that a rational constraint applies in the same way to everyone, it exhibits the impartiality that, we may suppose, identifies it with morality. Confirmation of this last claim is found if there is a significant overlap between the constraints that would result from the agreement of rational persons and the traditional principles of morality.

I find the argument condensed in the last two paragraphs extremely compelling. And it leads us into the heart of the contractarian understanding of morals and politics. Although we should not suppose that our actual moral practices and social institutions result from agreement, we may nevertheless hold that the appropriate justificatory test for the principles, practices, and institutions that govern and structure human interaction in ways that constrain the individuals involved is whether they would have been accepted by those individuals were they fully rational persons, each concerned to advance his own good (or the realization of his substantive aims), and collectively able to determine ex ante their terms and conditions of interaction by voluntary and unanimous agreement. This test is seen by the contractarian as affording a justification that is at once rational and moral. I defend a specific interpretation of it in *MbA*, claiming that fully rational persons would reach agreement by bargaining on the basis of the principle of minimax relative concession (MRC), and that their agreement would be voluntary and their subsequent interaction fully cooperative provided that none of

them bargained from an initial position that violated a revised Lockean Proviso, which essentially prohibits persons from bettering their own position by worsening that of others.

How far my particular account of the contractarian test affords an adequate way of specifying the underlying idea of rational agreement is evidently a matter for debate. I am now convinced that MRC needs at least some modification. Marlies Klemisch-Ahlert has offered a rigorous formulation of the intuitive ideas underlying the principle.[2] As she shows, for the general (*n*-person) case, MRC should be replaced by a lexicographic principle, which can be formulated as a maximin principle in relative utility gains (relative benefit in my terminology) as she does, or as a minimax principle in relative concessions. I believe that this meets the *formal* inadequacies of MRC for bargaining involving more than two persons. There may be, however, a deeper inadequacy, since both MRC and Klemisch-Ahlert's lexicographic principle ignore the structure of the interactions by which what is distributed in bargaining has been produced. Whether and how to refine either the formulation of MRC, or the way in which it is applied to interactions in order to accommodate this structure, are unresolved issues. My brief discussion in "Moral Artifice," where I am primarily concerned to reply to Jean Hampton's objections to MRC, only sketches some of the dimensions of the problem.[3]

To discuss the other part of the contractarian test – the revised Lockean Proviso – would take me far beyond the bounds of these brief remarks. The formulation of the proviso in *MbA* is at best only a first approximation.[4] But a more adequate formulation would leave unanswered the strong challenges that may be brought against its relevance for both agreement and compliance. Occasionally, I find myself tempted to discharge the proviso from its contractarian employment, giving it a strong letter of recommendation to rights theorists seeking a principled basis for their position. But then I find myself once more convinced that it is exactly that power of the proviso – to convert, as it were, a Hobbesian state of nature into a Lockean one – that is needed in a full contractarian moral theory. Hoping to clarify these matters in my unwritten paper (tentatively titled "Rational Choice and the Lockean Proviso"), I pass on to the principal theme of these comments.

Important as the contractarian test is, I have come to realize that it is not the central theme of *MbA*. Morality, as I have said, involves *constraint*,

[2] See Wulf Gaertner and Marlies Klemisch-Ahlert, "Gauthier's Approach to Distributive Justice and Other Bargaining Solutions," Chapter 11 in this volume.

[3] See Jean Hampton, "Equalizing Concessions in the Pursuit of Justice: A Discussion of Gauthier's Bargaining Solution," Chapter 10 in this volume, and my "Moral Artifice," *Canadian Journal of Philosophy* 18 (1988): 395–8.

[4] See the critique by Don Hubin and Mark B. Lambeth in "Providing for Rights," Chapter 8 in this volume.

but constraint is not in itself a moral concept. It is, however, the key concept of my theory, and it should be understood prior to introducing morality. This is not best done by relating it solely to the basis afforded by the PD for rational agreement. In *MbA*, I do, of course, recognize that to show the rationality of agreeing to mutual constraints is not to show the rationality of actually complying with or adhering to those constraints. I follow the familiar progression of Hobbes's argument in *Leviathan*, recognizing that agreement (the theme of his second law of nature) is distinct from compliance (the theme of his third law), and that the Foole's objections to the third law must be answered. But this progression obscures as much as it reveals, and I have come to the view that whatever may be the links between agreement and compliance (and I remain convinced that there are such links), the fundamental argument for the rationality of accepting and complying with certain constraints on the direct pursuit of one's substantive aims does not depend on supposing either that others will comply with them, or that it would be rational to accept them if and only if others agree to accept them as well.

The argument for the rationality of compliance is best understood if we recognize that it has no essential connection with the PD, and is simply an instance of a more general argument for the rationality of accepting certain constraints. And this more general argument has no direct connection with morality; the constraints it justifies need be neither mutual nor impartial. Consider a story quite different from that of the prisoners. You and I have been engaged in some clandestine activity, and each of us is now considering whether to reveal our involvement. Each of us knows that if either of us talks, he will represent the other's involvement unfavorably. Therefore, neither of us wants the other to talk. But if the other does talk, neither of us wants to have remained silent. I don't want my involvement revealed; I strongly prefer that we both be silent. You, however, would like best to tell your story, if you could do so before I tell mine. The situation has this structure:

	You talk	You remain silent
I talk	mutual second worst	my second best, your worst
I remain silent	my worst, your best	my best, your second best

You have a dominant strategy: talk. If I know your preferences, I know this, and I have a best response: talk. The outcome is, as in the PD, the mutual second worst. But if I could count on you to remain silent, then my best response would be to remain silent, and we should both do better. If you know my preferences, you know this, and so you want to convince me that you will remain silent – even if the price of convincing me is the cost of actually remaining silent. To be sure, you would prefer to deceive me. But you may believe, and be correct to believe, that your prospect of deceiving me is slim. Unless you are sincerely committed to remaining silent, I shall not trust you and shall talk.

Suppose you believe that, if you are committed to remaining silent, you will be able to communicate your commitment to me, so that there is a high probability that I shall remain silent. And you believe that if you are not committed to remaining silent, then there is a high probability that I shall talk. Suppose you prefer a high probability of mutual silence with the alternative that I alone talk to a high probability of mutual talking with the alternative that you alone talk. Then it is rational for you to commit yourself to the constraint of remaining silent. And this commitment is independent of any expectation you have about my willingness to accept any constraints.

In *MbA* I defend *conditional compliance* in PD-type situations. I argue that if you are confronted with a PD, it is rational for you agree to act in a way that is mutually advantageous rather than individually maximal, and to comply with this agreement given that you expect the other person to comply (or, in the case of several others, given that you expect sufficient compliance). It might seem, from this argument, that only conditional compliance is rational. It might be supposed, therefore, that constraint, to be rational, must be mutual. By treating morality as a matter of mutual constraint, it might seem that all rational constraint is moral. But the example we have just considered shows that this would be mistaken. In some situations, it is rational for an individual to accept a unilateral constraint on his efforts to maximize the realization of his aims.

I emphasize this, not only to remove the temptation to equate rational constraint with morality, but also to correct possible misinterpretations of my defense of conditional compliance in *MbA*. Some persons want to argue that it is rational to act in a mutually advantageous way in PD situations, provided others do so as well, because one gains more from their constraint than one loses from one's own. This is a bad argument, which I totally reject. However, in focusing on mutual constraint in *MbA*, my rejection of it may be less clear than it should be. It is rational to act in a mutually advantageous way in PD situations, if one gains more from one's own disposition to constraint, than one loses from one's actual exercise of constraint. This is a good argument; it is the argument of *MbA;* and it has nothing to do with mutuality.

I cannot embark here on the task of constructing a theory of rational constraint. I explore some relevant ideas in a recent paper.[5] The key idea might be expressed as *opportunity maximization*. The rational individual

[5] See "In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)," *Proceedings of the Aristotelian Society* 89 (1989): 179–94.

develops those dispositions, forms those intentions, and adopts those plans that afford him the most favorable opportunities, even though they may require him to act in ways that do not maximize the realization of his aims. In his interactions, others are led, through knowledge or belief about his dispositions, intentions, and plans, to behave in ways that enable him to do better, even though he does not maximize, than he could expect to do were he believed to be disposed to perform only maximizing acts. Thus, in our example, believing that you intend to keep silent, I keep silent; you then do better, keeping silent, than you could expect to do were I, believing that you would always maximize and so talk, myself to talk.

Prominent among the issues that a theory of rational constraint must face is the way in which constraint is to be characterized. Some, influenced by the view that the agent's aims are revealed in her choices, would insist that one cannot, literally, choose a nonmaximizing act. Wishing to assure me of your silence, you do not commit yourself to an act that would not best realize your aims, but rather you change your aims, coming to prefer silence to talk. Instead of treating your aims simply as exogenously given, we allow them to be endogenously revisable. You exhibit what, from the standpoint of your original aims, would seem to be constraint, but in fact you act so as best to fulfill your revised aims.[6]

Others, willing to admit that an agent can choose a nonmaximizing act, nevertheless deny that such a choice can be truly rational. Wishing to assure me of your silence, you commit yourself, quite rationally, to an act that itself is irrational. Constraint is an instance of rational irrationality.[7]

I reject both of these interpretations, holding that an agent can choose a nonmaximizing act and can do so with full rationality. You choose to keep silent, and your reason for so choosing is found in your belief that, had you not been committed so to choose, you would be less favorably placed and would be realizing your aims to a lesser extent than you are actually doing. But I cannot debate the possible interpretations of constraint here. I mention them simply to emphasize their importance on the agenda set by the idea of rational constraint.

In a revised presentation of the ideas of *MbA*, I should establish the rationality of certain forms of constraint before introducing the PD as revealing a structure of interaction calling for mutual constraints. And I should then focus on the interpretation of mutuality. In *MbA*, I treat mutuality as requiring *conditional* cooperation; a rational person should

[6] The position mentioned in this paragraph is suggested by Edward F. McClennen in "Constrained Maximization and Resolute Choice," *The New Social Contract*, pp. 108–18.

[7] The position mentioned in this paragraph is suggested by Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), pp. 1–23.

constrain her behavior in interactions with persons who, she expects, will exercise constraint themselves. Peter Danielson has argued that I should treat mutuality as requiring *reciprocal* cooperation; a rational person should constrain her behavior in interactions with persons who, she expects, exercise constraint in interaction with, and only with, those who, they expect, exercise constraint.[8] I take the issues raised by Danielson to be important and unresolved. His book, *Artificial Morality: How to Make Morality Rational*, will be a major contribution to advancing our understanding of the form that mutual constraint should take.

Only after addressing mutuality, would I introduce morality in my hypothetically rewritten *MbA*. I should appeal to the structure of the PD to show how the general argument for the rationality of certain constraints may be applied to the subclass of mutual constraints. And I should then characterize morality as a set of dispositions, both practical and affective, that enables agents who have the capacity for constraint to exercise that capacity so that they may engage in cooperative ventures for mutual advantage (to adapt, as I so often do, Rawls' useful phrase). In this way, we may reach a contractarian understanding of the *form* of morality. Note that this understanding makes no explicit reference to agreement; the contractarian element is provided entirely by the requirement of mutuality. And it leaves the content of morality quite unspecified; the dispositions that are considered moral virtues and the principles that establish moral obligations have yet to be rationally determined.

Is there a problem in determining them? It may seem that we need only consider which of those constraints that are generally accepted an individual should rationally follow. If morality consists in constraints that are both rational and mutual, what else could be relevant? Here is the point in a contractarian moral theory at which the idea of agreement is needed. As I noted near the beginning of these comments, in recognizing the benefits of mutual constraints, we are naturally led to consider them as possible objects of rational agreement. Each person must find it rational to agree with her fellows to constrain her behavior in certain determinate ways, provided she may expect others to do so as well. And it is the constraints that individuals would find it rational to agree to, rather than the ones that they actually find it rational to accept, that provide the content of contractarian morality.

The actual principles and practices that most persons accept and comply with need have little to do with realizing mutual benefit. Given that most persons accept them, and expect others to do so as well, a particular

[8] See "Closing the Compliance Dilemma: How It's Rational to be Moral in a Lamarckian World," Chapter 16 in this volume, and "The Visible Hand of Morality," *Canadian Journal of Philosophy* 18 (1988): 376–83. I discuss this matter, briefly and quite inadequately, in "Moral Artifice," *Canadian Journal of Philosophy* 18 (1988): 399–402.

individual may find it rational to accept them herself. Now from a descriptive standpoint, we may say that these principles and practices constitute the morality of those who follow them. But a morality in this sense need have no rational foundation or justification. The only rationale for complying with it is that such compliance is expected. And, of course, once this is recognized, the morality is undermined. Only if the principles and practices would be agreed to, by persons seeking the fullest realization of their aims, would there be any substantive rationale for compliance. Recognition of such a rationale sustains morality, and the constraints it imposes.

Agreement thus enters the contractarian characterization of morality, bringing rationality into the specification of its contents, so that we may determine the particular principles, practices, and institutions with which each person would find it rational to commit herself to comply, could she expect general compliance. Here then is the place for the argument that I sketched at the very outset of these comments in support of the contractarian test, as a standard for both the morality and the rationality of constraint. But note that this argument idealizes the appeal to agreement, by treating the contractarian test, not as a matter of what persons would agree to given their actual social circumstances, but rather as a matter of what persons would agree to were they to choose ex ante their social circumstances. I defend this idealization by arguing that moral structures that satisfy the contractarian test may be expected to have a stability lacking in principles and practices that depend for their rationale on the particular structure of an actual society.[9] The test appeals to an agreement among persons for whom the usual social contingencies of bargaining are altogether irrelevant, and who lack any power to coerce each other. Thus, in reaching such an agreement, each may exercise only his own rationality, addressed to the equal rationality of his fellows. The constraints that the test endorses are, therefore, rationally salient. This salience both facilitates convergence on them as a basis for interaction and inhibits departures from that basis.

The conclusion of the contractarian argument is not that rational persons, whatever their actual circumstances, must comply with the constraints of morality. No argument could show that morality is rational in that sense. What the contractarian argument does show is that rational persons will recognize a role for constraints, both unilateral and mutual, in their choices and decisions, that rational persons would agree ex ante on certain mutual constraints were they able to do so, and that rational persons will frequently comply with those mutual constraints in their interactions. I claim no more in *MbA*, and I remain convinced that I am entitled to claim no less.

[9] See "Why Contractarianism?" Chapter 2 in this volume, and "Morality, Rational Choice, and Semantic Representation: A Reply to My Critics," *The New Social Contract*, pp. 177–90.

Bibliography

Note: Only works closely related to Gauthier's work are included here.

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

Baier, Annette. "Pilgrim's Progress." *Canadian Journal of Philosophy* 18 (1988): 315–30.

Baier, Kurt. *The Moral Point of View: A Rational Basis of Ethics*. New York: Random House, 1965.

Baier, Kurt. "Rationality, Value and Preference." *Social Philosophy and Policy* 5 (1988): 17–45.

Barnett, Philip M. "Rational Behavior in Bargaining Situations." *Noûs* 17 (1983): 621–36.

Braybrooke, David. "Inequalities Not Conceded Yet: A Rejoinder to Gauthier's Reply." *Dialogue* (Canada) 21 (1982): 445–8.

Braybrooke, David. "The Maximum Claims of Gauthier's Bargainers: Are the Fixed Social Inequalities Acceptable?" *Dialogue* (Canada) 21 (1982): 411–29.

Braybrooke, David. "Social Contract Theory's Fanciest Flight." *Ethics* 97 (1987): 750–64.

Buchanan, James M. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago: The University of Chicago Press, 1975.

Buchanan, James M. "The Gauthier Enterprise." *Social Philosophy and Policy* 5 (1988): 75–94.

Campbell, Richmond, ed. *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press, 1985.

Campbell, Richmond. "Gauthier's Theory of Morals by Agreement." *Philosophical Quarterly* 38 (1988): 343–64.

Copp, David, and David Zimmerman, eds. *Morality, Reason, and Truth: New Essays on the Foundations of Ethics*. Totowa, NJ: Rowman and Allanheld, 1984.

Danielson, Peter. "The Visible Hand of Morality." *The Canadian Journal of Philosophy* 18 (1988): 357–84.

Danielson, Peter. *Artifical Morality*. London: Routledge, in press.