ROBERT C. KOONS

GAUTHIER AND THE RATIONALITY OF JUSTICE

(Received in revised form 29 January 1992)

David Gauthier's *Morals by Agreement* breaks new ground in the application of game theory to the ancient question of the rationality of being constrained by a disposition to be just.[1] Gauthier's success in this enterprise has been called into question by a number of commentators, including Jody S. Kraus and Jules L. Coleman.[2] In this paper, I would like to do three things. First, in section 1, I would like to propose a friendly amendment of Gauthier's definition of the disposition of 'narrow compliance', an amendment that resolves the problem discovered by Kraus and Coleman. Second, in section 2, I point out that the problem of defining fairness can be seen as essentially a coordination problem. I use this model to criticize Gauthier's defense of a Lockean state of nature as the appropriate baseline. In section 3, I point out a new problem, having to do with the generalization of two-person bargains to the n-person case. Specifically, I want to suggest that the possibility of coalition-building among sub-groups of the entire population raises new problems about the rationality of justice.

Gauthier's strategy for establishing the rationality of justice depends, first of all, on the assumption that the decisionmaking machinery in humans is plastic. One can choose to alter one's own decisionmaking procedures. In particular, one can choose to become a constrained maximizer, a decision maker who maximizes her preference-satisfaction (measured by a utility function) subject to certain exogenous constraints. The effect of these constraints could also be approximated (although only approximated) by effecting an alteration in one's utility function which assigns maximal disutility to the intentional violation of one of the constraints. Given the plasticity of decisionmaking procedures, it makes sense to raise the question: under what circumstances, if any, and with what conception of 'justice', would it be rational to change one's

Philosophical Studies **76:** 1–26, 1994.

decisionmaking machinery into one which respects the constraints of justice?[3]

Gauthier answers this question by, first stipulating that the circumstance must meet two conditions: (1) universal injustice results in a state in which everyone is worse off than he would be under conditions of universal justice, and (2) each of the participants must be, in respect of the nature of her decisionmaking machinery, at least translucent to the other participants, that is, each must have a reasonably good chance of identifying the decision-theoretic type of any other. Second, Gauthier defines 'justice' in terms of a distribution of the net benefits of cooperation which maximizes the minimum relative benefit to each of the participants, using a noncoercive state of nature as the base point for the comparisons.

Gauthier claims that if prevailing circumstances meet these two conditions, and justice is conceived of in this way, then it is rational for any agent to undertake a modification of her decisionmaking procedures that incorporates into them the constraint of justice. Gauthier's central argument for this claim is this: given the condition of mutual translucency, the probability that I will be taken to be a constrained maximizer is greater if I really am a constrained maximizer. People who are taken to be constrained maximizers have greater opportunities to benefit from the cooperation of others, since no rational agent exercises forbearance except in exchange for like forbearance from others, and no agent expects such forbearance from another unless she believes that that other is a constrained maximizer. If mutual forbearance results in a state in which everyone is better off than in a state of universal injustice, then these opportunities for cooperation are valuable. Therefore, maximizers who are constrained by justice enjoy greater opportunities to benefit from this cooperation.[4]

This argument establishes that a disposition to exercise forbearance in order to participate in schemes of mutual benefit may be a disposition which it is rational to adopt. It establishes that the virtue of *trustworthiness* in reciprocal exchanges is rationally desirable; it has not yet established that the virtue of *justice* is desirable, where justice is the disposition to afford others, and to demand for oneself, a fair share of the benefits of such cooperation. In order to establish the rationality

of justice, Gauthier argues first for the rationality of a disposition he calls the disposition for 'narrow compliance', and then Gauthier uses this result to establish the rationality of justice. An agent is disposed to *narrow compliance* if he is willing to participate in schemes of cooperation only if they afford himself a share which is at least fair (I will set aside for the moment Gauthier's arguments for his particular definition of 'fair'). If one is trustworthy but not narrowly compliant, Gauthier labels one's condition that of being disposed to 'broad compliance'. A broadly compliant agent permits himself to participate in any cooperative scheme which affords himself some net benefit, as compared to the state of universal noncooperation.

The argument for the rationality of choosing the disposition of narrow compliance proceeds as follows. Suppose, for contradiction, that it is rational to be broadly compliant. Then it would be rational for other agents to become 'less-than-narrowly compliant', that is, to be disposed to cooperate only if one obtains a lion's share, an unfairly large share, of the resulting benefit. Faced with less-than-narrowly compliant fellows, a broadly compliant agent will consistently be forced to accept a mere pittance as a result of cooperation. By contrast, a narrowly compliant agent does not invite such exploitation. A society of narrow compilers is in a state of equilibrium: no one has any incentive to become less-than-narrowly compliant, and so everyone enjoys a significant share of the benefits of cooperation. In fact, given the universally of narrow compliance, each agent has a compelling reason not to be less-than-narrowly compliant, that is, each agent has a compelling reason to be just.[5]

1. KRAUS AND COLEMAN'S PROBLEM, AND A SOLUTION

As Kraus and Coleman point out, Gauthier's argument for narrow compliance proceeds too quickly.[6] They point out that the rationality of switching from broad to narrow compliance depends upon one's expectations about the future dispositions of other agents, and, consequently, it is almost never rational to choose to be narrowly compliant. Suppose first of all that we live in a society in which almost no one is narrowly

compliant. If everyone is broadly compliant, then, as Gauthier argued, it may be rational for some agents to become less-than-narrowly compliant. The less-than-narrowly compliant (or the 'unjust', for short) forego some opportunities, since they are unable to cooperate with one another, but they are able to extract additional benefits from cooperation with the broadly compliant. If there are too many unjust, some of the unjust will have an incentive to shift to broad compliance, thereby gaining new opportunities for cooperation. An equilibrium state will eventually be reached in which the ratio of broadly compliant and unjust gives no one any incentive to shift in either direction. However, under no circumstances would it be rational for anyone to become narrowly compliant. An isolated, narrowly compliant person would forego all opportunities to cooperate with the unjust, while having only a negligible effect on the number of unjust persons.

Alternatively, suppose that we live in a society in which nearly everyone is narrowly compliant. If there are any unjust persons in the society, then each narrowly compliant person will have some incentive to become broadly compliant. A switch of one agent from narrow to broad compliance will increase the number of that agent's opportunities for cooperation while having little or no effect on the total number of unjust agents. A series of such defections from narrow to broad compliance could eventually transform this society into one in which almost no one is narrowly compliant. Alternatively, the society could reach a critical point at which one more defection from narrow compliance would dramatically increase the number of unjust agents. Only at this point would it be rational to stick with narrow compliance. The equilibrium point which corresponds to this point is not, contra Gauthier, one of universal narrow compliance. Hence, the universal rationality of justice is not established.

What has gone wrong with Gauthier's argument? The basic problem is this: the broadly compliant are *free riders*, who benefit from the deterrence of injustice provided by the presence of narrowly compliant agents without bearing any of the costs. In fact, the situation described by Kraus and Coleman has the game-theoretic structure of a prisoner's dilemma, where 'cooperation' corresponds with narrow compliance, and 'defec-

tion' corresponds with broad compliance. We are all better off if we are all narrowly compliant than if we are all broadly compliant, since narrow compliance deters injustice, but we are individually better off being broadly compliant, whatever the other agents do, since a broadly compliant agent enjoys a wider range of opportunities (i.e., opportunities to benefit from cooperation with the unjust-but-trustworthy).

	Narrow	Broad
Narrow	2,2	0,3
Broad	3,0	1,1

This prisoner's dilemma game is a fragment of a larger, n-person game, in which each person has three choices: unjust, narrow, and broad. In order to capture the essence of this game, the injustice-deterrence game, we need to consider at least three players. Representing this game requires the use of three 3×3 matrices, one for each of the choices of the third player. The first payoff is for the row-player, the second for the column-player, and the third for the matrix-player.

The 3×3 game G_1:

I. Player 3 chooses Unjust

	Unjust	Narrow	Broad
Unjust	0,0,0	0,0,0	4,1,4
Narrow	0,0,0	2,2,0	2,3,4
Broad	1,4,4	3,2,4	3,3,6

II. Player 3 chooses Narrow

	Unjust	Narrow	Broad
Unjust	0,0,0	0,2,2	4,3,2
Narrow	2,0,2	5,5,5	5,5,5
Broad	3,4,2	5,5,5	5,5,5

III. Player 3 chooses Broad

	Unjust	Narrow	Broad
Unjust	4,4,1	4,2,3	6,3,3
Narrow	2,4,3	5,5,5	5,5,5
Broad	3,6,3	5,5,5	5,5,5

The values in the matrices were derived in the following way. First of all, a value of 0 represents a state in which the agent is excluded from all cooperation. This occurs when an unjust agent confronts only narrow or unjust agents, or when a narrow agent confronts only unjust ones. The value of 1 represents the value enjoyed by a broad agent when cooperating with two unjust agents, who take the lion's share of the resulting benefit. The value 2 represents the value enjoyed by a narrow agent who succeeds in cooperating with only one other agent. The value 3 represents the utility of a broad agent who cooperates with one just (broad or narrow) and one unjust agent. The value 4 represents the utility of an unjust agent who succeeds in cooperating with only one agent. The value 5 represents the value of a fair cooperation among all three agents. Finally, the value 6 represents the utility of an unjust agent who is able to exploit both of the other agents (who are broadly compliant).

In this game, the choice of broad compliance weakly dominates that of narrow compliance: one can sometimes do better, and one can never do worse, by choosing broad over narrow. Suppose that we eliminate player 3's dominated choice (narrow) and omit his payoffs. Then we get the following two 2-person games.

A. Player 3 is unjust

	Unjust	Narrow	Broad
Unjust	0,0	0,0	4,1
Narrow	0,0	2,2	2,3
Broad	1,4	3,2	3,3

B. Player 3 is broad

	Unjust	Narrow	Broad
Unjust	4,4	4,2	6,3
Narrow	2,4	5,5	5,5
Broad	3,6	5,5	5,5

Game A has two equilibria in pure strategies: Broad-Unjust and Unjust-Broad. There is also an equilibrium in mixed strategies, in which each agent chooses Broad with probability 1/2 and unjust with probability 1/2. The expected payoff of this equilibrium is 2 for each agent. There is no equilibrium point involving the choice of Narrow.

Game B has two equilibria in pure strategies: Unjust-Unjust and Narrow-Narrow. The first equilibrium is quite strong, and hence stable: given the choices of the other two players, each of the three players has positive incentive to stick with their choices. The Narrow-Narrow equilibrium, by contrast, is weak, and therefore unstable. Given that player 2 has chosen Narrow and player 3 has chosen Broad, player 1 is indifferent between Broad and Narrow. If there is the slightest chance that player 3 might accidentally play Unjust instead, player 1 has an incentive to shift from Narrow to Broad.[7]

The 3-person game G_1 has a third equilibrium in pure strategies: the point Narrow-Narrow-Narrow. This is the point which Gauthier believed to be uniquely rational. Unfortunately, this equilibrium point is quite unstable. Each of the three players has no incentive *not to shift* from Narrow to Broad. If any agent thinks that there is the slightest chance that one or both of the other agents might shift from Narrow to Unjust, it is rationally obligatory for that agent to shift from Narrow to Broad. Additionally, if player 3 believes that one or both of the other players has shifted from Narrow to Broad, player 3 will shift from Narrow to Unjust.

In order to extract a prisoner's dilemma game from this sort of interaction, we must consider a 4-person game, G_2. To simplify the exposition, I will assume that players 1 and 2 cannot play Unjust, and players 3 and 4 cannot play Narrow. In addition, I will assume that players 1 and 2 are interchangeable in terms of their effects on 3 and 4. Rather than

presenting the entire game matrix, I will extract various fragments from it. First of all, for sub-game C I assume that players 1 and 2 both play Narrow (I list only the payoffs of player 3, the row-player, and player 4, the column-player.)

C. Players 1 and 2 play Narrow

	Unjust	Broad
Unjust	0,0	1,1
Broad	2,3	3,2

In sub-game C, the choice of Broad strictly dominates that of Unjust for player 3. Given player 3's choice of Broad, player 4 must choose Unjust. Thus, if players 1 and 2 play Narrow, 4 will play Unjust, and 3 will not. In sub-game D, I assume that either player 1 plays Narrow and 2 plays Broad, or vice versa.

D. One player plays Narrow; the other Broad

	Unjust	Broad
Unjust	1,1	3,3
Broad	2,5	4,4

In sub-game D, player 3 will play Broad, since that choice still dominates his choice of Unjust. Given the choice by player 3, player 4 must choose Unjust. Thus, if only one of the first two players plays Narrow, player 3 will still play Broad and 4 will still play Unjust. In sub-game E, I assume that both player 1 and player 2 choose Broad.

E. Players 1 and 2 play Broad

	Unjust	Broad
Unjust	5,5	6,3
Broad	3,6	4,4

In sub-game E, both players 3 and 4 will play Unjust, since that choice strictly dominates Broad. Given this analysis of sub-games C, D, and E, we can now turn our attention to the choices of players 1 and 2. If both players play Narrow, then player 3 plays Broad and player 4 plays Unjust. Players 1 and 2 are able to cooperate on fair terms with 3, and no cooperation occurs between 1, 2 and 4. If both players 1 and 2 play Broad, then players 3 and 4 both play Unjust, and maximal cooperation occurs, but on terms which are unfair to players 1 and 2. If one player plays Narrow and the other Broad, then player 3 plays Broad and player 4 plays Unjust. The player playing Broad benefits from cooperation with player 4, while the player playing Narrow foregoes this benefit.

F. Resulting fragment of the 4×4 game G_2 involving players 1 and 2

	Narrow	Broad
Narrow	2,2	3,0
Broad	0,3	1,1

In sub-game F, it never pays to play Narrow. The game G_2 ineluctably moves to the equilibrium in which players 1 and 2 play Broad and players 3 and 4 play Unjust, despite the fact that this outcome is worse for everyone than the outcome in which 1 and 2 play Narrow and 3 and 4 play Broad.

To solve this problem, the definition of Narrow compliance must be altered in such a way as to eliminate the free-rider problem. Not only should unjust agents be punished by being excluded from cooperation, so too should broadly compliant freeriders be punished. This idea has also been proposed by Peter Danielson.[8] The following is a revised definition, a definition of 'recursively narrow compliance':[9]

> An agent x has a disposition for recursively narrow compliance (RNC) iff x enters into a cooperative scheme S with agents $y_1, \ldots, y_n$ only if (1) S provides x with at least a fair share of the resulting benefits, and (2) every one of $y_1, \ldots, y_n$ has a disposition for RNC.

This sort of circular definition is called a 'recursive definition'. The circularity is not vicious. To demonstrate this, I will instead define a series of RNC dispositions: RNC_0, RNC_1, RNC_2, etc. RNC_0 is simply narrow compliance as defined by Gauthier. For each i greater than 0, we can define RNC_{i+1} as:

> An agent has the RNC_{i+1} disposition if and only if x enters into a cooperative scheme S with agents $y_1, y_2, \ldots, y_n$ only if (1) S provides x with at least a fair share of the resulting benefits, and (2) every one of $y_1, \ldots, y_n$ has the RNC_i disposition.

Finally, we can stipulate that an agent has the RNC disposition simpliciter if and only if she has the RNC_i disposition for every number i. To demonstrate how this modified definition solves our problem, I will turn again to the 3× game G_1. Let's understand the choice of Narrow to correspond now to choosing the RNC disposition. An agent who chooses RNC is unable to cooperate not only with the Unjust but also with the Broad. Equivalently, Broad agents are unable to cooperate with Narrow agents. The resulting modifications of G_1, namely G_3, has the following structure.

The 3×3 game G_3 (with RNC):

I. Player 3 chooses unjust

	Unjust	Narrow	Broad
Unjust	0,0,0	0,0,0	4,1,4
Narrow	0,0,0	2,2,0	0,2,2
Broad	1,4,4	2,0,4	3,3,6

II. Player 3 chooses Narrow

	Unjust	Narrow	Broad
Unjust	0,0,0	0,2,2	4,2,0
Narrow	2,0,2	5,5,5	**2,0,2**
Broad	2,4,0	**0,2,2**	4,4,0

III. Player 3 chooses Broad

	Unjust	Narrow	Broad
Unjust	4,4,1	4,0,2	6,3,3
Narrow	0,4,2	**2,2,0**	0,4,4
Broad	3,6,3	4,0,4	5,5,5

The outcomes whose values have changed from G_1 to G_3 have their entries underlined. The three changes which are crucial to understanding the difference between G_1 and G_3 have their entries also in boldface. There are four equilibria in G_3 in pure strategies: UBU, BUU, UUB, and NNN. All four of these equilibria are stable: each agent has an incentive not to deviate from the equilibrium, so long as she assumes that the other players won't deviate. The crucial difference between G_1 and G_3 is that, in G_1 the NNN equilibrium was unstable. In fact, in G_1 the choice of Narrow was weakly dominated by Broad, making it very likely that players would deviate from the NNN equilibrium. In game G_3, deviations from the NNN equilibrium in favor of Broad are punished by the other two players, since players with the RNC disposition refuse to cooperate with Broad agents.

Similarly, if the 4×4 game G_2 is modified in the same way, replacing simple narrowness by RNC, the resulting sub-game involving players 1 and 2 has the following structure:

	Narrow	Broad
Narrow	3,3	0,1
Broad	1,0	2,2

Suppose players 1 and 2 both choose Narrow. Next, consider whether it would be rational for one of them, say player 2, to shift from Narrow to Broad. By shifting to Broad, player 2 would not change the fact that player 3 is deterred from choosing Unjust, and player 2 would gain the opportunity of engaging in beneficial cooperation with unjust player 4. However, player 2 must now pay a price for this shift: she must sacrifice the opportunity of cooperating with player 1, who has RNC

and therefore refuses to cooperate with non-narrow agents. Since player 2 must sacrifice cooperation on fair terms with player 1 in exchange for cooperation on unfairly unfavorable terms with player 4, we may reasonably assume that the shift would involve a net loss for player 2. Therefore, this sub-game is no longer a prisoner's dilemma game; instead, it has the structure of a game of pure coordination. Both of the pure equilibria, NN and BB, are stable. Thus, the 4×4 game has two pure equilibrium points: one in which players 1 and 2 both play Narrow and neither 3 nor 4 play Unjust, and one in which players 1 and 2 both play Broad and both 3 and 4 play Broad. Clearly, it is rational for players 1 and 2 to coordinate their choices by both choosing Narrow, since this is preferred by both players to the BB equilibrium.

Danielson has argued in favor of what he calls "reciprocal cooperation", which corresponds to RNC_1 above.[10] An agent with RNC_1 cooperates only with those who cooperate only on at-least fair terms. Thus, an agent with RNC_1 cooperates only with those who have the RNC_0 disposition, which corresponds to Gauthier's simple notion of narrow compliance. An agent with RNC_1 refuses to cooperate with those who are broadly compliant, as well as those who are more-than-narrowly compliant, but is willing to cooperate with those who have RNC_0, despite the fact that members of this last group are themselves willing to participate with the broadly compliant.

Certainly, there is a potential cost to be paid by shifting from RNC_1 to the fully recursive RNC. Agents with RNC forego opportunities to participate with all agents who are not similarly disposed. Thus, agents with RNC cannot cooperate with agents with RNC_0 or RNC_1. However, if enough people in a society have RNC, then everyone has an incentive to adopt and to retain the RNC disposition. Anything short of RNC, like Danielson's reciprocal cooperation, does not have this crucial stability feature. Agents with RNC_1 allow agents with RNC_0 to be free riders. The RNC_1 agents suppress the presence of broadly compliant agents, which in turn is crucial to suppressing the presence of unjust agents. RNC_0 agents benefit from this deterrence of broad compliance without paying any of the cost. Thus, a society in which nearly everyone has RNC_1 is unstable: agents have an incentive, first, to defect to RNC_0 and, then, to defect to broad compliance, and, finally, to become unjust.

2. THE DEFINITION OF JUSTICE AS A COORDINATION PROBLEM

In the last section, I assumed that there was a unique unproblematic definition of fairness which could be incorporated into the definition of recursively narrow compliance. Of course, this cannot be assumed. Instead of a single RNC disposition, there are in fact infinitely many, each one corresponding to a different conception of fairness. In the last section, I argued that we have good reason to think that any one of these dispositions, if chosen universally, would result in a stable (and presumably Pareto-optimal) equilibrium. But how can it be determined which of the possible RNC disposition will or should be chosen?

The choice among competing conceptions of fairness is essentially a coordination problem. For simplicity's sake, let's assume that there are only two possible conceptions of fairness, and two corresponding forms of RNC, N1 and N2. In a simple society of two persons, the structure of the coordination game would be something like this:

	N1	N2
N1	3,2	0,0
N2	0,0	2,3

This is not a game of pure coordination, since there is also an element of conflict of interest. Player 1 prefers N1 to N2, since on the first conception of fairness, he receives a relatively large share. Player 2, in contrast, prefers N2 to N1 for the same reason. How do such conflicts resolve themselves?

One very critical factor is that of saliency. If there are a large number of possible coordination schemes, then rational agents are likely to fix upon the one which has the greatest saliency, as Thomas Schelling has argued.[11] For example, if you and I are trying to meet somewhere in New York City, but we forgot to fix a precise place and time, we would probably try to find one another at an especially salient time and place, like Times Square at noon. If there is such a uniquely salient coordination point, then we can restructure our decision matrix as a choice between the salient point and a random selection of non-salient points.

Since choosing the salient point greatly increases the probability that we will succeed in coordinating, that choice will strictly dominate the other.

One important form of saliency is that of historical precedent. To use Hume's example, if the rowers on a boat are trying to select one among the infinity of possible coordination patterns in their rowing, the most salient one will be the pattern they have already established by trial and error. Thus, once established, a pattern of coordination tends to persist. This clearly has application to the problem of selecting a conception of fairness. A society will tend to maintain its historically established conception, no matter how arbitrary it may appear sub specie aeternitas.

Still, it may be possible to say something about such coordination points in general, abstracting from a particular society. Presumably, a viable conception of fairness must be relatively simple and relatively easy to apply, given the limited information available. A society's conception of fairness must satisfy two independent requirements: it must be simple, and it must constitute a stable equilibrium point. Gauthier's argument for minimax relative benefit as the correct conception of fairness can be understood as the claim that only this conception can satisfy these two requirements simultaneously.

For a conception of fairness to constitute an equilibrium point for a society, it must be the case that when everyone purports to have the RNC disposition which incorporates this standard of fairness, no trustworthy (constrained) agent has any incentive to resign. (It is not problematic if some unconstrained agents resign, since their doing so will only improve the scheme from the point of view of the constrained, and, in fact, unconstrained agents will never resign, since they are trying to reap the benefits of cooperation without bearing any of the costs.) Gauthier is in effect claiming a conceptually simple coordination scheme can achieve this stability only to approximating the standard of maximin relative benefit. The degree to which untrustworthy agents actually benefit from the scheme depends on two factors: the benefit which the scheme assigns to them, and the cost to the trustworthy which is imposed by undetected unconstrained maximizers. The impact of the second factor on particular agents is difficult to predict with any precision. Consequently, a given

conception of fairness is likely to achieve stability only if it maximizes the minimum benefit which it assigns to any agent. The lower the benefit which the scheme assigns to any agent, the greater the probability that that agent will find it advantageous to resign from the scheme, given the actual costs to that agent of cheating by unconstrained maximizers.

Once we have agreed that maximin relative benefit is an essential feature of a feasible stable coordination point, we still must ask: maximin benefit relative to what? That is, we must address the problem of defining the base point and the unit of measurement. At first glance, it appears that the answer must be: the base point is the utility which an agent would achieve as a free agent, absent any cooperative agreement, and the unit of measurement is provided by the difference between this base point and the maximum utility which the agent could reasonably expect. The baseline is the utility which the agent would enjoy in an anarchic, Hobbesian state of nature in relation to the rest of society. That Hobbesian condition is the state which the agent would enter if she withdrew from the coordination scheme. Gauthier contends, however, that the correct base point is one corresponding to a noncoercive, nonexploitative Lockean state of nature.

Gauthier's contention could be based on the following sort of argument. If society distributes benefits with respect to a Hobbesian base point, then it would encourage individual agents to engage in predatory behavior, in order to improve their standing at that base point. This predatory behavior would necessitate costly defensive measures. Consequently, the society would end up in a state which is not Pareto-optimal: everyone could be better of by foregoing this pointless cycle of predation and defense.[12] This argument can be illustrated by the following game, G4. Each player has four choices: to adopt an RNC disposition incorporating a Lockean base point and to engage in predation, to adopt such a disposition without engaging in predation, to adopt a disposition incorporating a Hobbesian base point and to engage in predation, and to adopt such a disposition without predation. If the agents choose incompatible dispositions, coordination will fail to occur. Consequently, we can focus exclusively on the cases in which they choose compatible dispositions, as summarized in the following two matrices.

H.

	NL + P	NL + NP
NL + P	2,2	2,2
NL + NP	2,2	3,3

J.

	NH + P	NH + NP
NH + P	2,2	4,1
NH + NP	1,4	3,3

In sub-game H, in which both players adopt a Lockean base point, predation achieves no improvement in relative final standing. Since predation is costly, both to the predator and his victim, the choice of no predation dominates. In sub-game J, predation improves one's final standing to a degree sufficient to outweigh its cost. Hence, predation dominates, and society faces another prisoner's dilemma, ending up at the sub-optimal equilibrium in the upper left corner. Since the equilibrium in sub-game H is preferred by all agents to the equilibrium in sub-game J, rational agents will coordinate their choices by adopting a Lockean base point.

This argument is superficially plausible, but it possesses a crucial flaw. We assume that if society adopts a Hobbesian base point, the base point must correspond to some actual condition of society prior to the agreement on a social contract. In fact, it is quite possible for a society to adopt a purely hypothetical Hobbesian base point. A hypothetical Hobbesian state can be calculated in various ways, depending on the extent to which this state is allowed to differ from the actual state. An investment-sensitive Hobbesian baseline would correspond to the utility an agent could gain in a state of nature, given the actual distribution of predatory and defensive investments. An investment-insensitive Hobbesian baseline, by contrast, would be the utility an agent would gain in a hypothetical state of nature which is constructed by deleting from the actual state all investments in predatory and defensive resources, while retaining each agent's native, raw talents, abilities, and vulnerabilities. A society can achieve coordination on that basis without anyone having any incentive to engage in any actual predation or to invest resources in

acquiring or improving the means of predation and defense. It is only the hypothetical gains one would achieve, given one's raw, native abilities, in such a Hobbesian state of nature which affect one's relative standing. Thus, the original argument for a hypothetical Hobbesian base point as the one characterizing any stable solution stands, unrebutted.

There is another line of reasoning which the would-be defender of the Lockean base point might pursue. Suppose, for simplicity's sake, that we have a society of three persons. Player 1 is relatively strong (in the sense that she would do relatively well in a Hobbesian state of nature) but relatively unproductive. Players 2 and 3 are relatively weak but relatively productive. A conception of fairness which incorporates a Hobbesian state of nature might result in a state in which players 2 and 3 are worse off than they would be in the absence of player 1. In such a situation, surely it would be rational for players 2 and 3 to pool their strengths and eliminate player 1, either by killing her or, more humanely, by confining her in such a way that she poses no further threat. Suppose then, that players 2 and 3 form a coalition and engage in negotiations with player 1 as a unit. The following game results, with player 1 as row and the coalition of players 2 and 3 as column:

Game G5:

	NL	NH
NL	2,3,3	0,2,2
NH	0,2,2	3,1,1

Clearly, the coalition of 2 and 3 is now in a position to insist upon the Lockean base point, and player 1 has no choice but to accede.

This line of reasoning does not provide a justification for the Lockean base point, however. What it really suggests is that no agent in an n-person bargaining situation can rationally expect to advance a claim which is incompatible with the other agent's doing at least as well as they would in the n-1 person contract which would result from excluding him. It gives us reason to modify, not the baseline of the comparison, but the claim point, which provides the unit of interpersonal comparison. It suggests quite a new approach to the n-person social contract problem,

one in which a fair and rational outcome is defined recursively, beginning with the two-person case, and proceeding step-by-step to sub-bargains involving larger and larger numbers of persons. Thus, the base case of the recursion will be the use of the rule of maximin relative benefit to fix the rational outcome of a two-person bargain, defining relative benefit by comparison with a Hobbesian state of nature. Given the definition of rational outcomes for all possible social-contract bargains involving n-1 persons, the rational outcome of a bargain involving any n persons can be defined as one which maximizes the relative benefit (in comparison with the Hobbesian base point and a claim point consisting of the agents' maximum permitted claims). An agent's maximum permitted claim is the highest utility which can feasibly be provided to that agent, subject to constraint that all other agents receive at least as much utility as they would in the n-1 person social contract that results from excluding the agent.

In fact, in later papers, Gauthier has explicitly adopted such a conception of the maximum permitted claim. In *Morals by Agreement*, Gauthier stated,

> Each person's claim is bounded to the extent of his participation in cooperative interaction. For if someone were to press a claim to what would be brought about by the cooperative interaction of others, then these others would prefer to exclude him from agreement.[13]

In two responses to critics, Gauthier has interpreted this statement to mean that, in societies of more than two persons, each agent is limited to claiming no more than would be compatible with giving the rest of society what they would receive, if they were to cooperate with one another but not with the agent in question.[14] This is a substantially smaller claim than the greatest utility which is compatible with giving everyone else the utility he or she would receive *in the state of nature.*

This new notion of maximum claim is clarified by the following definitions:

Definition. A *rational outcome* of an n-person bargain involving the n members of a set A is any outcome which maximizes the minimum size of the ratio $(u - u')/u''$, where u is the agent's maximum permitted

claim, relative to A, u' is the utility which the agent actually receives, and u'' is the utility which the agent would receive in the investment-insensitive Hobbesian state of nature.

Definition. The *maximum permitted claim* of an agent x, relative to set A of size n, is the highest feasible utility which can be provided to x by a social contract of the members of A, subject to the condition that the social contract provide every other member y a utility at least as high as y would receive in some rational outcome of a bargain involving the members of the set $A - \{x\}$.

I will argue that, as the number n grows larger and larger, the maximin rule and the precise definition of the state of nature grows less and less important. Consequently, disputes about whether the maximum rule should be replaced by the Harsanyi-Zeuthen-Nash solution, and about how the base point should be defined grow more and more irrelevant. This claim is based on the following assumptions. Economists standardly assume that as any factor of production becomes more abundant, its marginal productivity declines. Moreover, as any good becomes more abundant, its marginal utility to a given agent declines. Therefore, as more and more participants are added to the social contract, the extent to which the utilities of those who are already participants can be increased grows smaller and smaller. As the marginal productivity (in this sense) of each agent declines, average productivity approaches closer and closer to its from below.[15] Thus, the problem of distributing to each agent her maximal claim (limited by her marginal productivity) becomes less and less acute.

The problem of justice consists of two separable problems: the problem of commitment to cooperation and the problem of the division of the resulting surplus. The first could be thought of as the problem of market creation, since without commitment to cooperation, market exchanges are impossible. The second is the problem of market amelioration: the distribution of benefits in the absence of perfect competition, that is, in the presence of oligopoly and oligopsony. Both of these problems result in sub-optimal states, and both are present in the transition from

a state of nature to a scheme of cooperation, but, as the size of society increases, the second problem gradually disappears. For a large n, an n-person bargaining problem approximates the state of perfect competition, with each agent receiving close to her maximum permitted share, her marginal contribution to the benefits accruing from cooperation.

3. THE PROBLEM OF TYRANNICAL COALITIONS

Although the argument add the end of the last section does support the idea that the rationally justifiable claims of an agent are limited by her marginal productivity and thus are, to a large extent, independent of her standing in a Hobbesian state of nature, the argument also raises some disturbing questions about the viability of Gauthier's project. If the bargaining process is to be genuinely impartial, distributing (when possible) equal benefits to each, each person must participate *atomistically* in the n-person social contract bargain. Even in the modified conception developed in the last section, coalitions play only a hypothetical role, limiting the claims which can be advanced atomistically by each participant in the bargaining process. However, once the possibility of forming coalitions and bargaining as blocs is invited into the theory, it may not be a trivial matter to exorcise it later.

Consider again a simple three-person society. In such a society, each agent has essentially four choices: three choices which consist of forming a bargaining bloc of two with one of the other agents, and once choice which consists of entering the social-contract bargaining process solo. To simplify the discussion, let's suppose that the only possible two-agent bloc consists of players 1 and 2. Suppose that if player 1 and 2 form a coalition, they are able to enjoy a reasonably high utility of 1 through bilateral cooperation. Suppose that there are two optimal cooperative schemes, one in which all share equally (in terms of relative benefit), and one in which the lion's share of the benefit goes to players 1 and 2. Each player has two choices: to offer to cooperate on the basis of scheme C1, and to offer to cooperate on the basis of scheme C2. The game matrix is as follows.

K. Player 3 chooses C1.

	C1	C2
C1	3,3,1.5	1,0,1
C2	0,1,1	1,1,0

L. Player 3 chooses C2.

	C1	C2
C1	1,1,0	0,1,1
C2	1,0,1	2,2,2

If exactly two players succeed in cooperating on the basis of one of the two schemes, each receives a utility of 1 and the other player receives 0, since he does not benefit from cooperation. There are two equilibria in pure strategies: one in which everyone cooperates on the basis of C1, and one in which everyone cooperates on the basis of C2. Players 1 and 2 prefer C1, and player 3 prefers C2.

Suppose that in addition to C1 and C2, there are four additional schemes which are technically feasible. These schemes, C3 through C6, have the following payoffs:

C3: (1,1,4) C4: (1,4,1) C5: (4,1,1) C6: (5,5,0)

The base point for the bargaining process if (0,0,0), since each player receives a utility of 0 in the absence of cooperation. The claim point, as defined in the last section, consists of the point (4,4,4), since each agent can receive (in one of the schemes C3–C5) a utility of 4 while still providing the other agents with the utility (of 1) which they would receive in the corresponding two-person social contract. Therefore, it is scheme C2 which is fair, since in it each agent receives a utility halfway between their base point value and their maximum claim.

However, this result does not continue to hold if we assume instead that players 1 and 2 form a permanent bargaining bloc before entering into negotiations with player 3. If we treat players 1 and 2 as a unit,

then their baseline value becomes 1, since they can receive a utility of 1 by cooperating with each other in the absence of player 3, and their maximum claim becomes 5, since the only bloc which remains after players 1 and 2 are excluded is a bloc consisting only of player 3, and player 3 can receive only a utility of 0 in the absence of cooperation with players 1 and 2. Thanks to scheme C6, it is possible, with the cooperation of all three agents, for player 3 to receive 0 while players 1 and 2 each receive 5. Given this new base point and claim point, it is now scheme C1, and not C2, which is fair, since C1 maximizes the minimum relative benefit, as defined by the new base and claim points. We can say that scheme C1 is fair relative to the $\{\{1,2\},\{3\}\}$ partition of society (a partition which treats players 1 and 2 as an indivisible bloc), but not relative to the partition $\{\{1\},\{2\},\{3\}\}$.

Thus, it would appear to be rational for players 1 and 2 first to commit themselves to cooperating as a coalition and only then to enter into a social-contract bargain with player 3. The resulting "fair" and rational solution is preferred by both player 1 and player 2 to the state which would result if they acted atomistically in relation to possible cooperation with player 3. What we have, in effect, is a model for the rationality of a tyranny of the majority. A cohesive majority, acting as a coalition, is able to impose relatively unfavorable terms on the other rational agents in their society. If the coalition of players 1 and 2 is optimal for each player, that is, if neither player is able to improve her standing by switching to a coalition with player 3 instead, then the tyranny-of-the-majority solution is a stable one.

In societies of more than three players, there is also the possibility of compound coalition-building. Suppose, for example, that we have a society of four agents. It might be rational for players 1 and 2 first to form a two-player coalition, then to negotiate favorable terms with player 3 to form a three-player compound coalition, and finally to arrive at a still more favorable social contract with player 4. Compound coalitions are needed to explain the rationality of autocratic monarchies. The monarch forms a nuclear coalition with a small band of intimate supporters, who benefit from their proximity to the sovereign. This band then cooperates, from a position of relative strength, with a wider, penumbral group, the

army and nomenklatura. Finally, a coalition is formed which is able, by virtue of its near monopoly on force and communications, to achieve a very favorable social contract with the rest of society. I will follow Gauthier in calling such a system of compound coalitions a *coalitional infrastructure*.

The possibility of coalition-building introduces an element of instability into the social contract. It may be that there are no coalitional infrastructures which are optimal for each of their participants. The actual coalitional infrastructure would then suffer from an intrinsic instability: some members of the coalition would have an incentive to withdraw from an existing coalition and form a new coalition with outsiders.

This sort of phenomenon has been the subject of much study in the field of cooperative game theory. A central concept developed there is that of the 'core'.[16] An infrastructure belongs to the core of a game just in case it guarantees each agent the maximum utility which he can be guaranteed by any possible infrastructure. Stable infrastructures belong to the core of the social contract game.

Unfortunately, in many games, the core is empty. No infrastructure is stable in such games. In particular, the possibility of compound coalitions greatly decreases the chances that the social contract game has a core. For example, in the four-person game described above, it is likely that it would be rational for player 3 to try to form a two-person nuclear coalition with player 4 and then to try to entice one of player 1 or 2 into splitting up their coalition in exchange for a penumbral position in a new three-person compound coalition. This would be possible, since each of players 1 and 2 would prefer being in a penumbral position with a player 3+4 bloc to being the absolute outsider.

Since compound coalition (and any infrastructure not in the core) suffer from this inherent instability, their success would depend on steps being taken to suppress the formation of alternative compounds. The freedom of speech and association of outsiders and those in the extreme penumbra of the dominant coalition would have to be severely circumscribed. Oppression is costly: the state which results might well be Pareto sub-optimal. It is at least conceivable that everyone, even the

monarch and his cronies, might be better off in a society in which a ban on non-core coalitions is respected. If this is so, then it would be rational for agents to incorporate such a ban into their conception of fairness.

Alternatively, rational agents might first incorporate a ban on compound coalition-building into their conception of fairness. This ban would greatly simplify the social contract game and increase the chances that, in the new game, the core is not empty. In such a society, agents would organize simple, core coalitions, resulting in a stable infrastructure. In all probability, a core coalitional infrastructure would include a dominant coalition, containing some large segment of the society, most likely a majority. This majority would then be able to extract relatively more favorable terms for its members in bargaining with smaller coalitions of outsiders.

Therefore, even if a ban on unstable coalitional infrastructures is rational, it does not follow that a ban on majoritarian tyranny and the unequal treatment which results is equally rational. The case for the rationality of such a ban must be sought elsewhere. The best care for such a ban depends on the fact of human ignorance about the future, which in turn depends on the unpredictability of social change. A tyrannical coalition may be locally stable, in the sense that no member of that coalition has any incentive to leave it and try to create an alternative coalition, without being dynamically stable. As technological and demographic conditions change, what was once a locally stable coalition may become unstable. Since these changes are unpredictable, no member of a dominant coalition can be certain that she will always be a member of the in group. Each member of a tyrannical majority must consider the real possibility that she will one day find herself the victim of an alternative majority which excludes her. If agents are sufficiently risk-averse, and if they do not heavily discount future ills, they may find it rational to forego present opportunities for tyrannical advantages and accede to a universal ban on all coalition building. This would mean accepting arrangements which are fair on the Gautherian, atomistic model of rational bargaining.

Unfortunately, these conclusions are loaded with provisions and qualifications. Where a society is deeply divided along ethnic, linguistic, cultural, or religious grounds, and where one such group is numerically

or technically dominant, given the difficulty of cross-cultural coalition-building, members of a tyrannical coalition may reasonably conclude that the chances of their one day being the victims of an alternative tyranny are slight. This problem is compounded if these agents heavily discount ills which lie in the remote future, as they may reasonably do. Consequently, the most we can say is that a disposition for justice may, under favorable conditions, be a rational choice.

NOTES

[1] David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986).

[2] Jody S. Kraus and Jules L. Coleman, "Morality and the Theory of Rational Choice," *Ethics* 97 (1987): 715–749.

[3] In fact, even if one has doubts about Gauthier's assumption of human plasticity, recent work on reputation effects in such games as the iterated Prisoner's Dilemma and the chain-store paradox (see, for example, D. M. Kreps and R. Wilson, "Reputation and Imperfect Information," *Journal of Economic Theory* (1982) 27: 252–279) suggests that rational, unconstrained maximizers can act as if constrained, so long as their action is observed by other agents who believe that they *may* be constrained maximizers. In any case, however the constraint may be achieved, there seem to be strong theoretical and historical reasons for believing that it is quite possible for a society to exist in which (1) for the most part, agents act as if constrained by certain rules, and (2) some deviation from these constraints occurs.

[4] Gauthier, *Morals by Agreement*, pp. 170–177.

[5] Gauthier, *Morals by Agreement*, pp. 178–179, 226.

[6] Kraus and Coleman, "Morality and the Theory of Rational Choice," pp. 737–738.

[7] Here I am appealing to Selten's notion of a trembling-hand equilibrium. (R. Selten, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* (1975) 4: 25–55) An equilibrium is trembling-hand perfect if it remains rational for each given a small probability that the other players may deviate from the equilibrium point. The equilibrium Narrow-Narrow-Broad is not trembling-hand perfect.

[8] Peter Danielson, "The Visible Hand of Morality," *Canadian Journal of Philosophy* 18(1988): 357–384; see pp. 376–381.

[9] RNC corresponds to Danielson's SC, Selfsame Cooperation. Danielson, "The Visible Hand of Morality," pp. 378, 382–383.

[10] Danielson, "The Visible Hand of Morality," pp. 382–383.

[11] Thomas Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960), pp. 54–58, 83–118.

[12] See Gauthier, *Morals by Agreement*, pp. 194–197.

[13] Gauthier, *Morals by Agreement*, p. 134.

[14] David Gauthier, "Morality, Rational Choice, and Semantic Representation," *Social Philosophy and Policy* 5 (1987): 198–199; David Gauthier, "Moral Artifice," *Canadian Journal of Philosophy* 18 (1988): 397.

[15] I'm assuming that society remains in a state of relative scarcity of agents: that marginal productivity always exceeds average productivity.

[16] The concept of the core was developed by D. B. Gallies and Lloyd Shapley in 1953. See Martin Shubik, *Game Theory in the Social Sciences* (Cambridge, MA: MIT, 1982), p. 136.

Department of Philosophy
University of Texas at Austin
Austin, TX 78712-1180
USA