Eidos V, 1
June/Juin 1986

Critical Notice/Etudes critique

Gauthier on Rationality and Morality*

Peter Vallentyne, University of Western Ontario

1. Introduction

David Gauthier's book represents the culmination of his work over the last twenty years on the theory of rational choice and on contractarian moral theory. It is the most important book on contractarianism since Rawls' *A Theory of Justice*[1] and is mandatory reading for anyone specializing in contemporary moral theory.

Gauthier does two distinct, although closely related, things in his book: (1) he defends a theory of rational choice, and (2) he defends a contractarian theory of morality. The two are related, since on his view moral principles are rational and impartial constraints on the pursuit of self-interest.

---

*David Gauthier, *Morals By Agreement* (Oxford: Clarendon Press, 1986). Pp. x + 367.

2. The Basic Framework

Gauthier assumes a framework in which individuals are individuated by their preferences, capacities, and situations. One might also add beliefs to this list, but Gauthier assumes (as is common in economic theory) that individuals are fully informed, and so all individuals have the same beliefs.

He defends the 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 (being considered, complete, transitive, monotonic in prizes, and continuous) the instrumental 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, except the above formal properties.

In *parametric choice* -- that is, in choice situations in which the agent takes his/her environment as fixed -- a choice is rational if and only if it maximizes expected utility (i.e., is the most effective means for achieving one's goals). But what about *strategic choice*? What is it rational to choose in choice situations in which the agent recognizes that the outcome of choice depends in part on the choices of other rational agents? What determines whether an *agreement* among rational agents is a rational choice for each of them? A large part of Gauthier's book is addressed to this question.

3. Rational Agreement: The Initial Bargaining Position

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 which 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. Gauthier has a new and interesting account of both steps.

A common specification of the initial bargaining position is as the *non-cooperative outcome* ("the state of nature" outcome). This is the hypothetical outcome of the uncoordinated pursuit of self-interest.[2] If agreement is based on the non-cooperative outcome, then, although the benefits of cooperation over non-cooperation will be distributed among the parties, the distribution of the benefits and costs of non-cooperation will be untouched. Each person will end up with the net benefits that he/she would obtain from non-cooperation plus a portion of the benefits of not engaging in such behaviour.

Gauthier argues that a rational agreement cannot be based on the non-cooperative outcome as the initial bargaining position. A rational agreement must, Gauthier claims, provide a basis for rational compliance, and agreements based on the non-cooperative outcome do not do this. In particular, it is irrational, he argues, for those who would be net victims of non-cooperative interaction (i.e., those who would be worse off in the presence of the non-cooperative activities of others than they would be if left completely alone) to comply with agreements based on the non-cooperative outcome. Such agreements would perpetuate the benefits and costs of coercive activity even though such activity would no longer take place.

Gauthier claims that, in order for there to be a rational basis for all to comply with rational agreements, the initial bargaining position must be the hypothetical result of non-cooperative interaction *constrained by the Lockean proviso*, that

is, of non-cooperative interaction subject to the constraint that no one makes him/herself better off by making someone else worse off.[3] From the point of view of *morality*, it is plausible that the outcome of non-cooperative interaction constrained by the proviso is a more appropriate initial bargaining position than the non-cooperative outcome, but that is *not* Gauthier's claim. His claim is that from the point of view of *rationality* it is more appropriate.

I am unconvinced by Gauthier's defense of the relevance of the proviso for the theory of rational choice. The initial bargaining position must, it seems to me, reflect how people would fare if they were to opt out of society; and from a rational point of view, there is no reason to refrain from violating the proviso (bettering one's position by worsening the position of others) with respect to people who have opted out of society. The issue, however, is controversial, and given my space limitations I shall have to leave my claim undefended here. In any case, Gauthier's argument is well worth considering in detail.

4. Rational Agreement: The Bargaining Solution

The bargaining position is this: Given a set of feasible options, and a privileged feasible option that is the initial bargaining position, which option (or set of options) is the rational choice? 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), which claims that rational agreement would fix on an option that maximizes the product of each person's excess utility over the initial bargaining position. Thus, for each feasible option O one calculates the value [U1(O)-U1(0*)] x [U2(O)-(2(O*)]... x [Un(O)-Un(O*)], where O* is the initial bargaining position point, and Ui is person i's utility function. According to the Nash solution, a rational agreement would maximize the above product.

Gauthier defends a different bargaining solution to the bargaining problem. 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. In symbols, the relative concession for person i, of option O is [Ui(Oi)-Ui(O)]/[Ui(O)-U(O*)], where O* is the initial bargaining position option, and Oi 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).

According to Gauthier's bargaining solution, then, rational agents would choose an option for which the highest relative concession is as low as possible. Very roughly, the justification for this solution is that the lower the greatest relative concession that is made, the less there are grounds for complaint from the person who makes the greatest concession, so rational agreement would minimize the best grounds of complaint. One's ground of complaint, according to Gauthier, is measured by one's relative concession, that is, by the proportion of the maximum admissible gain one could obtain from agreement that one gives up.[4]

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. The difference lies in Nash's acceptance, and Gauthier's rejection, of Condition Alpha (also known as the Independence of Irrelevant Alternatives, although this has nothing to do with the condition of the same name introduced by Kenneth Arrow

in the theory of social choice): an option that is a rational choice for a given set of feasible options and given initial bargaining position point, is also a rational choice for any *subset* of these options with the same initial bargaining position point. Gauthier rejects this condition because he holds that what is rational for one to accept depends on how favorable one's most favorable admissible option is. 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 rational choice given the same initial bargaining point and a *subset* of those feasible options. Indeed, Gauthier holds that in general such a option will *not* be a rational choice for the subset, if the subset was obtained by eliminating someone's most favorable admissible option.

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.

5. *The Rationality of Complying with Rational Agreements*

The combination of Gauthier's specification of the initial bargaining position (the hypothetical outcome of non-cooperatve interaction constrained by the proviso) and his bargaining solution (minimize the maximum relative concession) specifies what it is rational for agents to agree to. There remains, however, a further problem. Why should anyone *comply* with the terms of a rational agreement? More specifically, does rationality *always* require that we comply with the rational agreements. 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 it may in general be in one's self-interest to comply with agreements, at least sometimes, it seems, it is in one's interest not to comply.

Gauthier disagrees. He argues that under certain broadly characterized conditions rationality requires that fully informed, rational agents dispose themselves to comply with the terms of rational agreements. More specifically, he claims that 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 dispose ourselves to comply with such rational agreements when others are likewise disposed. For if our characters are sufficiently translucent, and we are not disposed to comply, we will be excluded from cooperative arrangements, because others will not trust us.

This is an extremely important argument, for it purports to show that there is a rational solution to the age old 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. All we need to do, Gauthier argues, is to properly understand the dictates of rationality.

Once again, I am unconvinced by the argument. I have argued elsewhere:[5] (1) that the argument fails even for fully informed, highly translucent, perfectly rational agents; and (2) even if the arguments succeed for such highly idealized agents, it still fails for real people who are only partially informed, only slightly translucent, and only moderately rational. Still, it's an important argument, and will surely receive a lot of attention in the literature.

6. The Conditions of Morality

So far we have been discussing Gauthier's views on rational choice. Let us now consider his views on morality.

Like Rawls and Hume, Gauthier takes the circumstances of morality (or more narrowly: justice), i.e., the circumstances under which moral questions arise, to be the possibil-

ity of mutual benefit from cooperation (collective and coordinated choice). Cooperation can be mutually beneficial, he claims, when and only when: (1) there is an awareness of *externalities* in the environment (i.e., people being affected without their consent by the actions of others, as when my neighbour's wild party wakes me up in the middle of the night), and (2) there is an awareness of a self-bias in human character. Let us consider each of these conditions in turn.

In a world in which there are no externalities, and in which the other conditions for a perfectly competitive market are satisfied, there are, Gauthier claims, no moral issues. Adam Smith's invisible hand argument (a formalized version of which has been proven mathematically by modern economists) establishes that in such a world the rational pursuit of self-interest will lead to Pareto optimal results (i.e., outcomes which are such that no alternative outcome makes someone better off without making some worse off). Results that are Pareto optimal leave no room for mutual benefit, since, by definition, one person can gain only by someone else losing. Thus, Gauthier claims, there are no moral issues.

Many object to this veiw, since it entails, that in a world in which there are no externalities, there is nothing wrong with refusing to help a person severely injured by a tree blown over by the wind, whom you happen to come across in the woods. The injury that the person is suffering is not the result of an externality (not the result of someone else's action); it is just his/her bad luck. Many people hold that it is wrong to refuse to give significant help to someone when one can do so at a relatively small cost -- even in a world of no externalities. Such people will find Gauthier's starting point for morality (and that of contractarian theorists generally) fundamentally misplaced.

Another way of making this criticism (or at least a closely related criticism) is that although the absence of Par-

eto optimality is sufficient to give rise to moral issues. it is not necessary. Pareto optimality is a very weak notion, compatible with one person being very well off and everyone else being very poorly off. Those with a less minimal vision of morality than Gauthier see the purpose of morality as not only guaranteeing Paeto optimality, but also choosing between different Pareto optimal results (for example, between a 100-1 split and a 99-99 split, when both are Pareto optimal). The presence of externalities is not a necessary condition for morality, it might be claimed, since the choice between different Pareto optimal results still remains in the absence of externalities.

What about the second condition that Gauthier claims is a necessary condition for moral issues to arise? What about the awareness of self-bias? If self-bias is understood as having a special interest in oneself (one's mind and body), the claim is highly questionable. For, suppose that I want only to maximize the number of trees in the world, and you want only to maximize the number of cows in the world. Neither you, nor I, takes any special interest in ourselves in any obvious sense. And yet, given that we have different goals, in some circumstances our goals will conflict (due to externalities), and so it seems that there is a moral issue as to how the conflict is to be resolved. So the awareness of self-bias is not, it seems, a necessary condition for morality.

The necessary condition is awareness of *different* goals (preferences). Self-biased preferences are but one way in which agents can have different goals. The well-known prisoner's dilemma illustrates that when there are externalities and *conflicting goals* each person pursuing his/her own goal can lead to an outcome that is not Pareto optimal. The prisoner's dilemma is usually formulated in terms of the narrow pursuit of self-interest, but the result remains as long as the agents have different goals. Gauthier is, I suspect, aware of this fact, but his discussion does not make it clear.

7. Gauthier's Contractarian Moral Theory

A contractarian ethical theory judges an action permissible if and only if it (or a joint action of which it is part; or rules of conduct to which it conforms) would be chosen by the members of society under specified condition. Gauthier's theory holds that an action is permissible if and only if it conforms to *rules of conduct* that would be agreed to by the rational individuals with whom the agent interacts if they are fully informed, purely self-interested, and fully aware of their preferences, capacities, and situations, and if the initial bargaining position were the hypothetical outcome of non-cooperative interaction constrained by the Lockean proviso.[6]

Unlike Rawls's theory, which imposes a thick veil of ignorance on the agents, and thereby reduces the choice of principles to the choice of a single individual, Gauthier's theory is a genuine contractarian theory in that it is based on rational negotiation among many fully informed, determinate individuals. For both theories the relevant agreement is hypothetical, but the relevant circumstances of agreement are grounded much more closely in reality on Gauthier's view than on Rawls'.

Like Rawls, Gauthier has us assume that the parties to the agreement take no interest in each other. I have argued elsewhere,[7] and can but mention here, that the rationale for this assumption is suspect. Why not use realistic assumptions that reflect people's actual preferences (which to some extent are other-regarding)? Of course, it is important to show that the *existence* of mutually acceptable rules of conduct in no way depends on anyone's being concerned for others, but surely the *content* of these rules should depend on people's actual preferences.

As noted earlier, taking the initial bargaining position to be the outcome of non-cooperation constrained by the proviso has the effect that rational agreements will allocate benefits

differentially to people on the basis of their differing capacities. Those with greater mental and physical capacities will come to the bargaining table with a greater share of the good things than those with less capacities. Although the benefits of cooperation will be shared, the net benefits of constrained non-cooperation are left untouched. Thus, Gauthier holds that people deserve the benefits of exercising their mental and physical capacities -- as long as they do not worsen the situation of others. Unlike Rawls, and most socialist thinkers, Gauthier does not view these capaities as a common asset for all members of society. The possessors of these capacities -- not soceity -- is entitled to the benefits that these capacities may bring. Gauthier agrees with Rawls that no one deserves the capacities he/she has (since they are basically determined by one's biology and the social circumstances of one's childhood, neither of which one has much control over), but he denies that this means that people do not deserve the benefits of whatever capacities they have. A person's capacities determine how one would fare in the absence of others, so surely, he argues, they are relevant for determining how one should fare in the presence of others.[8]

The issue of whether one deserves the benefits of exercising one's capacities is a central issue in contemporary moral philosophy. Gauthier's defense of a fairly strong claim of individual desert is an important contribution to the debate.

8. Archimedean Choice

Gauthier's main argument concerning morality, we have seen, is that under appropriate conditions: (1) Rationality dictates that we comply with the terms of rational agreements. (2) The requirement for such compliance is impartial. (3) There are no other impartial rational requirements. (4) morality consists of impartial requirements of rationality. (5) Therefore, morality requires that (and only that) we comply with the terms of rational agreements.

Gauthier purports to offer an independent, and more traditional, argument for his conclusion. He claims that a fully rational, fully informed, impartial, ideal agent would choose the principle of minimax relative concession based on the Lockean proviso as the fundamental constraint on conduct. The fact that this principle would be chosen by such an impartial chooser is supposed to provide an independent justification of the principle as a moral principle.

But wait a minute. As I have interpreted Gauthier, the basic moral principle is that under certain general conditions we should conform to whatever rules of conduct would be agreed to by the members of society. The principle of minimax relative concession based on the Lockean proviso determines (assuming that Gauthier is correct about rational agreement) what rational agents would agree to. To help legitimize Gauthier's view of morality, we would need an argument that an ideal chooser woul choose rules of conduct *on the basis of* minimax relative concession based on the Lockean proviso. This would ensure, assuming some sort of ideal chooser view of morality, that the rules of conduct that would be rationally chosen by the members of society are the correct moral rules of conduct.

Gauthier does provide an argument that an ideal actor would choose on the basis of minimax relative concession based on the proviso. Very briefly, he claims that the ideal chooser would reason on the basis of *conditions common to all actual agents.* Given Gauthier's view on rational agreement, it is but a short step to the claim that the ideal chooser would choose on the basis of the minimax principle. He goes on, however, to claim to show that the principle of minimax relative concession based on the Lockean proviso would be -- not merely the *basis* for choice, but also -- the *object* of choice. In choosing principles of interaction, the ideal actor would, Gauthier claims, choose the minimax principle.

I find this part of the argument very confusing. First, why do we need this part of the argument at all? If the purpose is to provide an independent argument for his claim that an action is permissible if and only if it conforms to rules of conduct that would be rationally chosen by the members of society, then what is needed is an argument that an ideal actor would choose rules of conduct on the *same basis* (i.e., on the basis of the minimax principle). As I indicated, Gauthier does give such an argument, but that is not the argument under discussion. The conclusion of the argument in question is that an ideal actor would choose *as an object of choice* the minimax principle.

One way of interpreting the argument, which I think is not Gauthier's is that the argument provided is not supposed to be an independent way of reaching the same conclusion, but rather an argument for a new conclusion. The new conclusion is that not only will the members of society and the ideal actor reason on the basis of the minimax principle, they will also choose it as the fundamental rule of conduct. That is, the new argument purports to establish that the rule of conduct that would be selected by the minimax principle is the minimax principle. The argument would be structurally like a claim by a rule utilitarian that the rule of conduct that would maximize social welfare is the act utilitarian principle.

So interpreted, I find the argument rather unconvincing. Given our limited mental capacities, it is very plausible that some set of more concrete rules (e.g., "Don't kill", etc.) would be much more effective in regulating human conduct than abstract principles such as act utilitarianism and the minimax principle. Even if at the bargaining table people are assumed to be fully informed and perfectly rational, in real life people are not so ideal. Since the rules of conduct are to apply to real life people, and they are not very good at processing rules, a more concrete set of rules will be more effective.

Consequently, it is highly plausible that neither act utilitarianism nor the minimax principle would maximize social welfare, or minimize the maximum relative concession.

Of course, the issue is sufficiently complex to warrant much more discussion. All I have done here is to suggest an alternative interpretation of the mysterious part of Gauthier's argument, and to suggest that it is unlikely to succeed.

9. Gauthier's Moral Methodology

Where there is an awareness of externalities and of our differing goals, and where our dispositions are sufficiently translucent, rationality, Gauthier argues, imposes an impartial constraint on the pursuit of self-interest, namely that we comply with rational agreements. Gauthier explicitly equates moral constraint with this rational and impartial constraint. Although most people would accept that the claim that being a rational and impartial constraint on conduct is a *necessary* condition for being a moral constraint, many would reject the claim that it is a *sufficient condition.* One might calim, for example, that an elementof 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.

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

> Trusting theory rather than intuition, we should advocate the view of social relationships sketched in this chapter [and the book in general] without regard to the intellectual fashions of the moment. 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. (p. 269)

Gauthier does, I think, two things in this passage. One is that he rejects any appeal to moral intuitions -- whether they be intuitions about the adequacy of a general moral principle or theory, or about very specific cases. He is only interested in questions of rationality, not those of morality -- except to the extent that they reduce to questions of rationality. The other thing he does is come out in favor of intuitions about theories, and against intuitions about particulars. The main implication of this second point is that in assessing his theory of *rational choice*, he accepts appeals to the intuitive adequacy of his abstract general principles, but rejects appeals to the intuitive adequacy of its implications for specific issues. Justification, he claims, goes from the general principle to the particular implications -- not vice versa. Consequently, he rejects the relevance of testing his theory of rational choice to see how well it captures our considered judgements in reflective equilibrium -- even if only judgements about rationality are allowed, and all moral judgements are excluded. The reflective equilibrium test recognizes the relevance of considered judgements about both specific cases and about general principles, and so is rejected by Gauthier.

Thus, because Gauthier rejects any critcisms that are based on appeals to moral intuitions (particular or general), his project is best understood as a radically reformist conception of morality. It is not merely that his theory fails to capture some (or even most) traditional moral concerns, but rather that its connection with these concerns is purely contingent. His real concern is with rationally acceptable norms of interaction.

10. Conclusion

In summary, Gauthier's book concerns both the theory of rational choice and the theory of moral choice. With respect to the former he defends a new specification of the initial bargaining position (non-cooperative interaction constrained by the Lockean proviso), a new solution to the bar-

gaining problem (minimax relative concession), and a defense of the claim that under certain circumstances rationality requires us to comply with the terms of rational agreements. With respect to the theory of moral choice he offers a tough-minded contractarian theory according to which an action is permissible if and only if it conforms to the rules of conduct that real life people, fully aware of their circumstances, capacities, and preferences would agree to.

There are, of course, many aspects of *Morals by Agreement* that I have not touched upon. To mention but a few: Gauthier has an important criticism of Harsanyi's defense of utilitarianism.[9] He defends at length his individualistic and economic approach to morality. And he investigates some implications of his theory concerning income tax rates, inheritance, economic inequality, and other issues.

As I hope is apparent, *Morals by Agreement* is full of solid argument for novel claims. It is well worth reading.

NOTES

1. John Rawls, *A Theory of Justice* (Cambridge, Mass.: Belknap Press of Harvard University Pess, 1971).
A close runner-up is James M. Buchanan, *The Limits of Liberty: Between Anarchy and Leviathan* (Chicago: University of Chicago Press, 1975).

2. See, for example, James Buchanan, *The Limits of Liberty*.

3. 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 s 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). Given greater space I would explore two issues: 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 (more Lockean?) version that prohibits bettering one's situation by worsening -- as means or as a side effect -- another's situation?

4. Neither Nash's solution nor Gauthier's require that utility be interpersonally comparable. Because both involve utility differences (e.g., Ui(Oi)-Ui(0*)) they do not require that the zero points of different people's utility scales be comparable. Because (unlke 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.

5. "Gauthier on the Rationality of Compliance", unpublished.

6. Given Gauthier's view that under certain conditions rationality requires one to comply with rational agreements, and his contractarian view that morality requires us to comply wtih rational agreements, it follows that under these conditions rationality requires one to be moral. Gauthier thus has an answer to the question "Why be moral?".

7. "Contractarianism and the Assumption of Mutual Unconcern", unpublished.

8. Note that Rawls, like Gauthier, uses an initial bargaining position (Rawls' state of nature) that accords differing benefits on the basis of differing capacities, but his thick veil of ignorance negates the effects of the inegalitarian starting point by blocking all information about how specific individuals would fare.

9. This also appears in "On the Refutation of Utlitarianism", in *The Limits of Utilitarianism*, by Harlan Miller and William Williams, eds. (Minneapolis: University of Minnesota Press, 1982).