

16 June 2003

**GAUTHIER, David Peter (1932– )**

David Gauthier was born in Toronto, Ontario, on 10 September 1932. He grew up in the city and took his B.A. at the University of Toronto in 1954. He studied at Harvard University where he received the A.M. in 1955 and at Oxford University where he received the B.Phil in 1957 and the D.Phil in 1961, the latter under the supervision of J.L. Austin. His main academic positions have been at the University of Toronto (1958–1980) and the University of Pittsburgh (1980–2001) where he was Distinguished Service Professor of Philosophy. He also held visiting professorships and research appointments at a number of institutions including UCLA, the University of California, Berkeley, Princeton University, the Australian National University, All Souls College, Oxford, and the Ecole Polytechnique (Paris). Gauthier became a Fellow of the Royal Society of Canada in 1979.

Gauthier's career has for the most part been confined to the academy. But he always has had a keen interest in politics and was active early in his career. He was executive member of various political and pressure groups, including the Toronto Committee for Disarmament, the Committee of Concern for South Africa, the Canadian Civil Liberties Association, and the Committee for an Independent Canada, and in 1962 was a candidate for election to the Canadian House of Commons.

Gauthier's main contributions to philosophy have been in ethics and moral theory, the theory of practical rationality and the formal theory of rational choice, political philosophy, and the interpretation of early modern moral and political philosophy (especially Hobbes and Rousseau). He was one of the principal theorists, with the philosopher John Rawls and the Nobel Laureate James Buchanan, responsible for the revival of contractarian theory, and he was one of the first philosophers to introduce decision and game theory to moral theory. He has written widely in political theory and on a number of topics in politics (secession, nuclear deterrence, democracy, public reason). He is also the author of a number of important essays on the work of different contemporary philosophers (for instance, K. Baier, G. Grant, J. Harsanyi, J. Rawls, T. Scanlon, A. Sen).

Gauthier's writings on these diverse topics are for the most part connected. From his earliest work he has been preoccupied by the question of the rationality of morality – what reasons do we have to be moral? – and his interest in this question frames his conception of ethics. Gauthier's identification of morality with principles and constraints, his conception of them as conditional on the compliance of others, and his account of their specific content are all shaped by his concern with the rationality of morals. The full title of his first book – parts of which are revisions of his D.Phil. dissertation – is *Practical Reasoning: The Structure and Foundations of Prudential and Moral Arguments and their Exemplification in Discourse*. Over several decades he developed a contractarian account of morality which sought to establish both the principles of morality and the rationality of acting in accordance with them. Teaching seventeenth and eighteenth century moral and political philosophy in the late fifties and early sixties, Gauthier came to appreciate Hobbes' thought, and the first of many writings on his work, *The Logic of Leviathan*, appeared in 1969. In the late sixties while a visiting professor at UCLA he was introduced by Howard Sobel to game theory and to the Prisoner's Dilemma. Hobbist moral and political theory and contemporary game theory have been important influences on Gauthier's thought, even if he has in recent decades moved away from many aspects associated with both traditions.

The most complete statement of his moral theory is given in *Morals by Agreement* published in 1986. He has modified his views in a number of respects since writing the book. Some of his earlier essays, collected in *Moral*

*Dealing* (1990), provide an easier and more accessible entry into his thought. And some of the modifications of his theory are introduced in later essays. The questions taken up in many of his writings on Hobbes and Hume are relevant to understanding his moral theory. Gauthier's interests in Rousseau, especially the latter's psychological and biographical writings, are not necessarily those one expects from the creator of morals by agreement. Several of his essays on Rousseau, along with some early essays, may be seen as exploratory self-critiques.

Gauthier's morals by agreements is a theory about the nature and rationality of morality. We may usefully think of the theory as having several parts or elements. The first is an account of the human condition – the aim of practical reason, the natural condition of humankind, the function of constraints on action. Next is a account of the principles of conduct that rational agents would hypothetically agree to – a kind of “social contract”. The third element is a revisionist account of practical rationality essential to the argument aiming to show that virtually everyone under normal circumstances has reason to accept and to abide by the constraints imposed by these principles. Lastly, Gauthier argues that the principles in question are principles of morality. Much of the interest of the theory turns on the details of the account, most of which are innovative and original. But it is important not to lose sight of the aim and general structure of the theory. Gauthier aims to understand morality and to ascertain our reasons to be moral, and this he proposes to do by an argument which takes us from a particular account of the human condition to a conception of moral principles emerging from hypothetical agreement to a conception of rationality according to which we have reason to be moral in the conditions in which we typically find ourselves. The structure of the overall argument as well as its ambitions are important in order to avoid some common misunderstandings of the theory.

Gauthier follows Hobbes, Hume, and many others in thinking that humans characteristically find themselves in “the circumstances of justice”. The phrase is from Rawls who borrows from Hume and H.L.A. Hart a summary account of the conditions in which humans typically benefit from cooperation. These circumstances consist principally in scarcity, relative to our needs and wants, and self-bias, our tendency to favor ourselves and those close to us over others. Cooperation in these circumstances is mutually beneficial, and it is made possible by our capacity to constrain our self-seeking behavior by adhering to just principles of action. In *Morals by Agreement* Gauthier uses the neo-classical economic theory of perfectly competitive markets to illustrate his conception of the rationale for moral constraints. In a perfectly competitive market – the highly idealized markets of certain branches of neo-classical theory – all agents are rational and self-interested, and all exchanges are mutually beneficial (or Pareto-improving). In these markets there is no way of rearranging things so as to improve the situation of some without making others worse off. (In technical terms, the outcome of perfect competition are a Nash equilibrium, Pareto-efficient, and in the core.) In such a world there is no place, no need, for mutually beneficially principles of action; individual rational choice, through trade, secure all of the benefits available. In real markets, of course, there are public goods and externalities (e.g., clean air, congestion) and many opportunities for mutually beneficial cooperation. But Gauthier's purpose in referring to perfectly competitive markets is first to give an example of a world, even if largely hypothetical, where there would be no need for a rational morality and secondly to have us think of moral principles as, in effect, designed to resolve externalities and other “market failures”. Given that markets typically presuppose the existence of a number of constraints on self-seeking behavior (for instance, a regime of property rights), it is not clear what are, if any, the policy implications of Gauthier's account of perfectly competitive interaction. In the book it mainly seems to illustrate the theoretical possibility of a “morally free zone”, one in which the constraints of morality would have no place, a kind of “moral anarchy”.

In *Morals by Agreement* Gauthier defends a subjectivist and instrumentalist conception of practical

rationality according to which we are rational to the extent that our acts maximize the satisfaction of our considered preferences. Rationality on this view is purely instrumental; no particular ends are rationally required. In his more recent thinking Gauthier has moved away from the subjectivism of this account. It is not in any case necessary for the conclusions that he wishes to defend about morals; the assumption he needs is that human reasons for action are characteristically agent-relative, namely that something's being a reason for me does not entail its being a reason for others. Gauthier's early subjectivist account of course entails this agent-relativity, but it is not necessary for it. The book presentation of the theory also supposes that agents considering the terms of interaction with others reason from the perspective of their own interests – the assumption of “non-tuism”. But this assumption also is not essential to the theory, and Gauthier has moved away from it.

Gauthier argues that agents in the circumstances of justice have reason to constrain their behavior by accepting and adhering to mutually advantageous and fair principles of action conditional on the compliance of others. To determine what they are he considers the principles that suitably characterized rational agents would choose were they to consider the question. Unlike Rawls and others, he imposes no “veil of ignorance” and does not deprive his hypothetical contractors of knowledge of who they are and where they find themselves. And, again unlike Rawls and others, he represents the hypothetical choice situation as a collective bargain. The principles selected represent a compromise, each person making a concession from their maximal claim. The principles selected govern the distribution of benefits and burdens amongst a set of cooperators; they determine the distribution of the cooperative or social surplus, the gains that cooperation make possible. So the talents and assets that are prior to or independent of social cooperation – our “natural assets” – are not subject to redistribution as they are in Rawls' theory. The fact that some bring more as it were to the bargaining table than others – they are more talented or fortunate – does not imply that others must be compensated for their lesser natural assets. Distributive justice is concerned with the cooperative or social surplus.

Gauthier relies on bargaining theory, part of the theory of games, to determine the principle of distribution that rational agents would select. He defends a principle of minimax relative concession which says that the maximum relative concession that anyone must make should be minimal, where relative concession is measured by an individual's maximal and minimal claims to the cooperative surplus. The principle says that no one may receive a relative benefit smaller than necessary. Under certain conditions it requires equal relative concessions. The principle does not require interpersonal comparisons of utility; rather it would have us look at an interpersonal comparison of the proportion each person's potential gain from cooperation that must be conceded. Gauthier's principle is similar to the solution to the two-person bargaining problem axiomatized by Ehud Kalai and Meir Smorodinsky. Most game theorists find the traditional solution developed by John Nash more plausible, and Gauthier himself has come to favor it over his earlier position. This change of mind undercuts the argument offered in *Morals by Agreement* for the principle of minimax relative concession, and it is unclear whether an argument for it can still be made.

Gauthier's principle of minimax relative concessions is in fact only one part of what I have identified as the second element of his contractarian theory. He also develops an account of the initial bargaining position, one which is highly original and which makes it clear that his theory is only in part a Hobbist one. Gauthier argues that rational agents interacting in a pre-moral state would come to accept certain principles and that these constrain the baseline or status quo point for the application of the distributive minimax relative concession principle. His account here is “Lockean” in appearance and has features similar to Robert Nozick's conception of basic rights. But Gauthier, inspired by James Buchanan's two-stage contractarian political theory developed in the latter's *Limits of Liberty*, constructs a conventionalist argument different from the natural law foundation of Lockean theorists. Gauthier argues

that rational agents in a pre-moral state would constrain their interactions by “a Lockean proviso” which prohibits bettering one’s situation through interaction that worsens the situation of another. Imagine that you come to a river and find a bridge which you may cross provided you pay the builder a small fee. Imagine also coming to a mountain pass which someone prohibits you from crossing without paying a small fee. Consider the options you would have faced had neither the bridge builder nor the mountain pass toll collector existed; in the first case there would be no bridge for you to use, but in the second the pass would still exist. These simple examples illustrate the counter-factual test offered by the proviso; the mountain pass toll collector violates the proviso in charging you a fee, but the bridge builder does not.

Gauthier uses the proviso to develop an argument for the emergence of limited rights (which I have dubbed “semi-natural” elsewhere) which protect each person’s exercise of their own powers. Thus individuals, prior to agreeing to be governed by minimax relative concession or other principle, acquire a right to their natural assets and in the fruits of their labor. These rights are limited in certain respects, but they constrain the application of minimax relative concession. The latter is applied to an initial bargaining situation in which agents have these rights. Gauthier’s account and argument is remarkable in that he provides in effect a prospective, conventionalist case for rights and duties that many have thought presupposes a natural law framework.

Gauthier’s moral theory presented thus far offers an account of the nature and the content of morality – what morality asks of us and why. It does not yet establish that we have reason to comply with the demands of morality. Justice often require in ways that we act contrary to our interests or aims – we may not steal, cheat, or break our word when this would prove advantageous to us or to the causes we defend. As many thinkers from Plato to Philippa Foot have noted, it often pays to be unjust. Gauthier’s response to this fact and the problems it poses for the kind of account he develops is to argue that we misconceive practical rationality, even instrumental rationality, if we think the aim of rationality determines in any straightforward way the manner in which we should reason or deliberate. The aim of rationality – say, to do as well as possible – does not necessarily determine our principle of decision – for instance, to choose the best alternative at each moment of choice. In terms of the utility-maximizing conception of rationality which he has accepted until recently, Gauthier argues that the aim of maximizing utility does not mean that we should, at each decision point, maximize utility. Instead we should reason in ways which are utility maximizing. Just as it is sometimes the case that we do best or at least well by not aiming to do best or well. So it may sometimes be that the utility maximizing course of action is not to maximize utility at each decision point. The point can be expressed and developed without presupposing a utility-maximizing or any particular account of rationality. Given that our mode of reasoning or deliberation itself affects our prospects, our aims or purposes are sometimes best served by our not seeking to do best at every decision point.

Gauthier’s discussion in *Morals by Agreement* is conducted in terms of “dispositions to choose” and specifically of “constrained maximization”, the disposition to cooperate with other cooperators even in circumstances where defecting is more advantageous. In later work Gauthier develops his revisionist account of practical rationality in terms of rational plans and intentions and of modes of deliberation. If we grant that agents may do better in any number of circumstances by acting in ways that are not “straightforwardly maximizing”, the problem is to determine how acting as a constrained maximizer is rational. In the book Gauthier assumes that if our dispositions to choose is rational, then our choices determined by these dispositions are also rational. A number of theorists have followed Thomas Schelling in arguing that it is often rational to do things that are irrational, but they argue that the latter do not in the circumstances cease being irrational. Gauthier thinks that if a course of action is better than any other in its effects, then it may under certain conditions be rational to adopt it and to intend to carry

out its element even if some of them are not, from the standpoint of the moment of execution, the best thing to do in terms of one's aims or purposes. He seeks therefore to establish that if a mode of deliberation or a plan of action is rational, then acting according to it can be rational even if so acting requires doing things that are not, considered from the standpoint of the moment of action, optimal.

Principled action constrains one's action, and it is rational to be so constrained. Thus, if Gauthier is right, it can be rational to abide by certain norms or principles, even when they require acting in ways that are not best from the standpoint of the time of action. Much of Gauthier's work since *Morals by Agreement* has sought to develop and to defend his revisionist account of practical rationality. Some of his later essays compare his account to those of Edward McClennen and Michael Bratman and address the arguments of critics.

If all Gauthier were to establish – and it would be no small feat – was that we have reason to accept and to abide by certain principles of cooperation, then he would not have shown that we have reason to be moral. The last element of his theory is an argument, interspersed throughout *Morals by Agreement*, identifying the various principles and dispositions as moral. His argument is in effect a functional one: the principles and dispositions in question resemble familiar moral ones in important respects. Impartiality seems to him to be a defining feature of morality, and it is central to the argument he makes for identifying the principles derived from rational interaction and bargaining with those of morality.

From an account of the human condition and the circumstances of justice, Gauthier develops a theory of the principles that rational agents would hypothetically agree to and a revisionist account of practical rationality which would establish the rationality of “principled” behavior. Lastly, he argues that these principles and dispositions to choose should be identified as moral. The theory has been the object of intense critical examination, and each part of the account has been criticized. Many contemporary moralists reject the starting point, the account of the circumstances of justice. They think that justice speaks even outside of the context of potential cooperation. On Gauthier's view there is no room for moral constraint outside of the context of mutual benefit. Even if one relaxes, as he now does, the assumption of mutual disinterest or non-tuism, there may be some apparent moral obligations that have no place in a Hobbist or Humean framework (for instance, duties to the infirm or unproductive, to non-human animals). And left egalitarians will remain dissatisfied with the restricted scope of distributive justice on Gauthier's account. Many philosophers and decision theorists have quarreled with aspects of Gauthier's contractarian derivation of principles, some favoring other principles, others the alternative Rawlsian conception of a hypothetical social contract. Gauthier's revisionist account of practical reason is widely viewed as implausible, both by thinkers influenced by decision-theoretic or economic conceptions of rational choice and by traditional philosophers. Lastly, some have quarreled with Gauthier's identification of principles of rational cooperation as moral. Some of these criticisms, especially the last kind, may rest on misunderstandings of the theory. But others do not. There are many details of the account that may not be right, as Gauthier himself has argued in later essays. The widespread rejection of the revisionist account of rationality is striking. It is also puzzling, as one might have thought that any moral theorist hoping to establish the rationality of moral action would need an account of reasons like Gauthier's that would allow for counter-preferential or principled choice. In this regard it is odd that the neo-Kantianism dominant in contemporary American thought has abandoned the attempt to establish the rationality of morals, seeking only to secure the accord of “reasonable” people.

Gauthier's other writings, especially those on important early modern thinkers (Hobbes, Locke, Hume, Rousseau, Kant), are important contributions to philosophy independently of the merits of his moral theory. They establish that his concerns, his aims, and to some extent his style, while contemporary in many respects, are also

similar to past thinkers. He says on the opening page of *Morals by Agreement*, “What theory of morals... can ever serve any useful purpose, unless it can show that all the duties it recommends are also truly endorsed in each individual’s reason?” This is a thought that presumably would elicit not only Hobbes’ assent, but also that of Plato, Kant, and many others.

## BIBLIOGRAPHY

*Practical Reasoning: The Structure and Foundations of Prudential and Moral Arguments and their Exemplification in Discourse* (Oxford, 1963).

“Morality and Advantage”, *Philosophical Review* 76 (1967): 460–75.

*The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes* (Oxford, 1969).

ed., *Morality and Rational Self-Interest* (Englewood Cliffs, N.J., 1970).

*Morals by Agreement* (Oxford, 1986).

*Moral Dealing: Contract, Ethics, and Reason* (Ithaca, N.Y., 1990).

“Commitment and Choice: An Essay on the Rationality of Plans”, in *Ethics, Rationality, and Economic Behaviour*, eds. F. Farina, F. Hahn, and S. Vannucci (Oxford, 1996), pp. 217–243.

“Assure and Threaten”, *Ethics* 104 (1994): 690–721.

“Public Reason”, *Social Philosophy & Policy* 12 (1994): 19–42.

“Intention and Deliberation”, in *Modeling Rationality, Morality, and Evolution*, ed. Peter Danielson (Oxford, 1998), pp. 41–54.

“Political Contractarianism”, *The Journal of Political Philosophy* 5 (1997): 132–148.

“Rethinking the Toxin Puzzle”, in *Rational Commitment and Social Justice: Essays for Gregory Kavka*, eds. Jules Coleman and Christopher W. Morris (Cambridge, 1998), pp. 47–58.

*The Social and the Solitary* (tentative title of a book on Rousseau, forthcoming with Cambridge University Press).

### *Further Reading*

Boucher, David and Paul Kelly, eds. *The Social Contract from Hobbes to Rawls* (London and New York, 1994).

Buchanan, James. *The Limits of Liberty* (Chicago, 1975).

Danielson, Peter. *Modeling Rationality, Morality, and Evolution* (Oxford, 1998).

Gauthier, David and Robert Sugden, eds. *Rationality, Justice, and the Social Contract: Themes from **Morals by Agreement*** (New York and London, 1993).

Kraus, Jody S. *The Limits of Hobbesian Contractarianism* (Cambridge, 1993).

Lessnoff, Michael, ed. *Social Contract Theory* (New York, 1990).

Morris, Christopher W. “A Contractarian Account of Moral Justification”, in *Moral Knowledge? New Readings in Moral Epistemology*, eds. Walter Sinnott-Armstrong and Mark Timmons (Oxford, 1996), pp. 215–242.

Morris, Christopher W. *The Social Contract Theorists: Critical Essays on Hobbes, Locke, and Rousseau* (Lanham, Md, 1999).

Morris, Christopher W. and Arthur Ripstein, eds. *Practical Rationality and Preference: Essays for David Gauthier* (Cambridge, 2001).

Paul, Ellen, et al., eds. *The New Social Contract: Essays on Gauthier* (Oxford: Blackwell, 1988).

Ridge, Michael, “Hobbesian Public Reason”, *Ethics* 108 (1998): 538–68.

Vallentyne, Peter, ed. *Contractarianism and Rational Choice* (Cambridge, 1991).

3,920 words

Christopher W. Morris  
University of Maryland, College Park  
[cwmorris@umd.edu](mailto:cwmorris@umd.edu)