Review for *Economics and Philosophy*

by Peter J. Hammond, Department of Economics
European University Institute, Badia Fiesolana, I-50016 S. Domenico (FI), ITALY;
and Stanford University, CA 94305-6072, U.S.A.

Morality within the Limits of Reason, RUSSELL HARDIN. Chicago: University of Chicago Press, 1988, xx + 234 pages.*

I. BOUNDED MORALITY

Much of economics presumes that economic agents can formulate and solve decision problems of unbounded and unimaginable complexity. Not only can they work out instantly how to play a perfect game of chess; even while blindfolded they can also play perfectly against all the other chess-players in the world simultaneously. Utilitarian morality has in the past often made similarly unrealistic assumptions about moral agents' capacities. This book represents a welcome departure from this tradition, though not the first, since similar ideas can certainly be found in the uncited work of Richard Hare (1981) and no doubt elsewhere as well.

The author states his aim clearly at the beginning of the preface (p. ix):

> "In what follows I wish to reconstruct utilitarianism on the basis of two fundamentally important classes of considerations that bear on the problem of choosing in social life: these are the limits of reason that prevent us from even approaching ideal conditions and the pervasive problem of strategic interaction in bringing about good results. Another major class of considerations is difficulties in the value theory of utilitarianism. I will address this class of difficulties at some length, but I will not pretend to offer a constructive resolution of these difficulties."

---

* The Book Review Editor Margaret Schabas suggested considerable improvements, and Russell Hardin himself made some sympathetic comments which helped to reduce even further the already quite small

As is claimed on p. xvi and elsewhere, much can be said about what are the likely "consequences in the form of human welfare" which result from any action, even without a complete value theory for assessing those consequences. Indeed, understanding what actions cause which consequences is a "structural problem" like knowing what constraints have to be faced. Value theory, on the other hand, is concerned with ethical objectives. Economists, of course, are used to separating objectives from constraints, even though the separation is not always kept as clear as it should be.

In the book, therefore, "morality" is taken to mean some form of reconstructed utilitarianism, though the value theory which underlies utilitarianism is acknowledged to be still incomplete. This is precisely because utilitarian value theory still presents difficulties in specifying appropriate ethical objectives. I shall return to this point in Section V below.

After the preface, the book consists of an "Introduction" followed by five chapters, divided into 37 numbered sections. This should help explain the references in the review which follows. The ordering of these section also has a compelling logic which I shall follow quite closely.

II. WITHIN THE LIMITS OF PRACTICAL REASON

Section 2 of the introductory chapter discusses in more detail some implications of what Herbert Simon has called "bounded rationality". Unfortunately, I have never found Simon's particular formulation of this theory all that convincing, based as it is on a form of imperfect maximization which he calls "satisficing". It is better, in my view, to recognize that agents often maximize very effectively among the small number of actions and their consequences which they typically consider when making a decision. The bounds on rationality arise because agents can only contemplate very few of the actions that are open to them in principle, and cannot work out many of the enormous number of possible consequences which could result. They are forced to use only small and unsophisticated decision models, in effect. Then, however, Hardin (p. 8) fails to separate the bounds due to imperfect reasoning as clearly as he should from those due

to lack of information, which is not really a "limit of reason" at all. After all, with unbounded rationality, a lack of information can be overcome by modelling as different states of the world all the possibilities that the moral agent may be unaware of. The same could also be said of the lack of relevant causal theories concerning the consequences of actions. Really, what matters in every case is that a moral agent can only be expected to have a very limited model of the true decision problem.

Section 5 shows how bounded rationality begins to make sense of the distinction between act and rule utilitarianism. David Lyons pointed out that the optimal rule is always to choose that act which maximizes utility, so that the distinction is vacuous. But, as Richard Hare argued, a simple rule makes fewer demands on agent's reasoning and modelling abilities than does the complicated rule of being an act utilitarian.

This introductory chapter concludes with an intelligent discussion in Section 7 of the role of those far-fetched or "peculiar" examples which are often used in attempts to refute utilitarianism or consequentialism. One example due to William James (p. 23) is of a bargain permitting utopia to be had for almost everybody at the cost of "a certain lost soul on the far-off edge of things [leading] a life of lonely torture." Hardin skirts dangerously close to the untenable claim that any ethical theory should only apply to a restricted domain of problems which excludes such far-fetched examples. He does, however, suggest that "a world that encompassed" such examples "would defy virtually all our understandings" (p. 27). Actually, if it were not for bounds on reason, it would surely be right to allow an unrestricted domain, as in Kenneth Arrow's theory of social choice. Otherwise the theory would have nothing to say about certain problems which may now seem remote, but could become urgent later on. Bounds on reason, however, imply bounds on the scope of any practical ethical theory. This seems the best argument for not giving undue weight to examples which are so remote from commonsense experience that any assessment of the sum of individuals' utilities is bound to be extremely untrustworthy.

III. WITHIN THE LIMITS OF CONTROL

The very brief Section 3 of the preliminary chapter raised the curtain on the problems of strategic interaction which game theory seeks to address, and how these constrain what moral agents can achieve. In my view it severely over-estimates the significance of "co-operative" game theory with transferable utility. Nor are it and the ensuing Chapter 2 at all attuned to late developments in game theory during the 1980's such as the notion of "rationalizable" strategies due to Douglas Bernheim (1984, 1986) and David Pearce (1984). These, however, are but minor complaints. Thereafter, Chapter 2 gets down to an extensive but elementary game-theoretic analysis of some important examples of strategic interactions. This is hardly profound, but like much of the other material in this book, it does cut through much confusion that can be found in earlier philosophical writings. The reader who is an economist or game theorist should be warned that the numbers in each example are preference rankings rather than payoffs or utilities. Thus lower numbers are preferred to higher. This leads to some confusion on p. 85, which mentions payoffs. Mixed strategies or probabilistic beliefs are never considered, so the numbers have only ordinal significance.

Two highlights of this part are the clear distinction between interests and preferences on p. 38. Also, Section 13 on "Promising" is especially sensible in recognizing how, along with honesty and truth telling, keeping promises is extremely important in those ethical environments possessing what economists would call "asymmetric information" — particularly when there is moral hazard.

IV. WITHIN THE LIMITS OF FEASIBLE INSTITUTIONS

Chapter 3 applies the previous game-theoretic discussion of Chapter 2 in order to consider those institutions or contractual arrangements which are Pareto efficient, in the appropriate sense that they cannot be replaced by any alternative feasible institutions whose equilibrium outcomes or consequences are Pareto superior. The value theory required is therefore very incomplete, since only changes whose consequences raise the welfare of everybody are approved. As remarked on p. 79, the topic has become

"institutional" rather than "act" or "rule" utilitarianism — although, as discussed on p. 101, this may be one concept of rule utilitarianism which John Rawls had in mind earlier.

Section 16 begins by discussing general issues raised by rights as an important instance of institutional arrangements. A key paragraph on pp. 78-79 describes rights as:

> "institutional devices for reducing the burden of gathering information and calculating consequences. They differ from rules . . . in their being backed by institutional force but they are similar to the rules of thumb of some rule-utilitarians in their actual function. Unlike rules of thumb . . ., however, such institutional rules as legally defined rights cannot easily be overridden when calculation shows that in a particular case a better outcome would follow from violating the rules … [T]he costs of setting up the devices of deciding on when to violate the rules are too great … When this conclusion seems not to follow in a particular case, then we may institutionally, as we do individually, resolve that case against the rules."

This view of rights as a way of economizing on information is very similar to one that has been offered by Partha Dasgupta in a paper cited much later, and in a totally different connection, on p. 136. Dasgupta, it is true, puts more emphasis on ignorance as a constraint that can never be violated, rather than as something which could be overcome at a cost which may, however, be too high to make it worthwhile. Both Dasgupta and Hardin, however, come close to my own favourite view. This is that when individuals have private information which affects the costs and benefits of particular actions, the outcomes have to be regarded as functions of that information. These "outcome functions" must also satisfy certain "incentive constraints" which arise because of strategic interaction in a game of incomplete information. Moreover, although sometimes private information can be uncovered and disclosure enforced, there are always costs to doing so which should be included among the consequences to be evaluated. Not many people like to live in a "police state."

A key advantage of this theory of rights is described on pp. 80–81:

> "If rights are viewed as metaphysical, abstract, or directly intuited rather than contingently derived, they must be defined very simply without many subclasses, distinctions, or exceptions. Legal rights, as opposed to moral or human rights, do not suffer from this disability in principle. They can be modified to meet new situations or conditions, and they can be highly articulated with manifold subclasses, distinctions, and exceptions."

After this preliminary discussion, the next three Sections (17–19) consider in turn "protections" at the individual, binary ["dyadic"] and group levels. Individual protections usually concern property rights, binary protections freedom of contract. As discussed on p. 91, protection of property rights is also viewed by many as useful at the group level, in ensuring "general utility" — presumably meaning Pareto efficient perfect market outcomes. However, it should be noted that this common view is unjustified when there is distributive injustice in the historicably determined allocation of property rights.

Section 20 notices the difficulties in relating observed behaviour in group decision problems to individual welfare — a point made elsewhere by Amartya Sen (1977) and many others. Finally, Section 21 gets to what perhaps should have been the main topic all along — namely, institutions as utilitarian constraints on poorly reasoned behaviour. Then Section 22 considers how the previous discussion of formal legal institutions can be extended to practices and rules such as promise-keeping. A difficulty arises in the interesting case when keeping a promise imposes costs on the promise-maker which have to be weighed against benefits to the promisee. It is argued (p. 107) that "because interpersonal comparisons of the relevant kind cannot be made with precision", the utilitarian advantage of breaking the promise has to be large and "convincing." At first, this seems very loose: if the best calculation I am capable of suggests that action A_1 is better than action A_2 — even only very slightly better — then how can it not be right for me to do A_1 rather than A_2? What is so special about keeping a promise that only allows me to break it when I am very sure of the benefits

(after taking into account all the incentive effects alluded to in Section 13, of course)? Why not apply the argument in reverse, and argue that a promise should always be broken unless the costs of doing so far outweigh the benefits of keeping it? Since promise-keeping can be seen as a particular kind of institutional rule, a much better argument would essentially be the same as that advanced on pp. 78-79 (as quoted above) for justifying institutional rules in general.

Chapter 3 concludes with some concise discussions of Sen's "liberal paradox" (Section 23) and of "contractarianism" (Section 24). Chapter 4 goes on to consider important issues such as beneficence and distributive justice which the Pareto principle cannot settle on its own, because the gains of some individuals have to compared with the losses of others. A number of points should be made. First, the discussion on pp. 116 and 126 only illustrates the distinction between distributive justice and beneficence, without giving a proper definition of either. It seems that distributive justice is intended as a "macro" concept which each individual's beneficence is powerless to affect. Second, when distributive justice is discussed in Section 27, on p. 127 it is supposed that only ordinal interpersonal comparisons can be made. This is also not really explained, though the examples discussed do make it fairly clear that Hardin has in mind something like the version of Patrick Suppes' (1966) justice criterion which is due to Amartya Sen (1970) — namely, a change is a "Suppes improvement" if, after permuting individuals and their welfare levels if necessary, there would be a Pareto improvement. More specifically, when the set of individuals is fixed, consequence x is better than consequence y if to every individual under y there corresponds a unique individual under x with a higher welfare level.

Many of the examples in Chapter 4 relate to economics — air pollution (p. 118), rights to subsistence (p. 124), Hayek's excuse for inequality (p. 129), the choice of an economic system (p. 131), the duty to alleviate poverty (p. 136), and the whole of Section 32 on "welfare, incentives, and policies" — with its fairly sophisticated discussions of usefulness of transfers in kind (cf. the analysis by Blackorby and Donaldson, 1988), and of the dangers of land reform in encouraging too many people

to remain dependent on agriculture for a livelihood. In between a good case is made for some instances of state paternalism (Section 28) when a particular kind of behaviour is institutionally utilitarian, and for "institutionalized interventions" (Section 29) to prevent practices such as duelling (p. 144), marriage or labour contracts in perpetuity (pp. 145-146), or workers agreeing to longer hours than might be thought desirable (p. 147).

V. WITHIN THE LIMITS OF WELFARE THEORY

As explained above, the first four out of five chapters (and 32 out of 37 numbered Sections) are almost exclusively devoted to exploring the constraints upon ethical decisions and the consequences which they can lead to. There is some presumption, of course, that human welfare should be promoted. In Chapter 3 the Pareto criterion is freely used, and Chapter 4 even discusses issues that involve some (weak) interpersonal comparisons. Chapter 5 (and the last five sections) are devoted to problems in completing "utilitarian value theory". As explained in the very last paragraph of the book (p. 207):

> "We can proceed on the analysis of structural issues without first constructing a fully adequate welfare theory. And we can proceed with the analysis of welfare theories, such as economic utility theory, independently of their use in utilitarianism."

In the ultimate sentence I find myself for the first time in rather serious disagreement with Hardin. An all-embracing ethical theory, after all, will have to be able to resolve such issues as population policy and other matters requiring preferences between different kinds of individual. Such a theory will remain incomplete unless it can handle an unrestricted domain of potential and hypothetical decision problems. If utilititarianism is to be a complete ethical theory in this sense, interpersonal comparisons of utility will surely be revealed because of the implied preferences for different kinds of people. As decisions with risky consequences need to be considered, there will even be a cardinal interpersonally comparable utility function with a determinate zero utility level. Ratios of utility levels must signify marginal rates of substitution between different kinds of person. This I have recently set out to explain in Hammond

(1990). In this way the notion of utility value becomes inseparably linked to the ethical preferences revealed by solutions to certain hypothetical ethical decision problems. Nevertheless, there is still the separable issue of how to construct the utility function on which all ethical decisions are supposed to be based. After all, the utility function is just a convenient mathematical construction showing the links between the solutions to all possible ethical decision problems. Nor do I wish to contradict the last sentence of the introduction to this chapter (p. 169):

> "The chief difference between Bentham's utilitarianism and any credible utilitarianism today is that no theory today can plausibly be based on additive, interpersonally comparable, cardinal utility without radical revision of the best present understanding of such a utility theory."

especially as this book was written well before its author could have been expected to react to the ideas expressed in this paragraph.

Of course, the utility theory I am contemplating is one which looks for the "chimera" (p. 178) of a best state of affairs — indeed, a best complete collection of contingent states of affairs, depending on what the theory of choice under uncertainty generally calls the "state of nature". This does indeed often make "impossible demands on our reason" (p. 197). It clearly needs replacing by a more complete theory allowing models of decision problems to be bounded. So the utilitarian "program" (p. 205) still needs refinement. But perhaps it does no harm to see first what one could do with "unbounded rationality" before recognizing our limitations.

VI. WITHIN THE LIMITS OF INFORMAL REASON

The author can hardly be blamed because his reason was too limited to be able to anticipate some developments in value theory after the book was written. Somewhat less excusable, however, is the inconclusive discussion in Chapters 3 and 4 of many important policy issues. I was frankly rather frustrated on several occasions to see the analysis end just as things seemed to be becoming really interesting. Of course, it must be admitted that the same is true of almost all the other philosophical works on ethics

which I have seen. An alternative procedure was established by Derek Parfit in *Reasons and Persons*, using simple examples. These, however, often fall into the class of those "peculiar examples" which Hardin is right to distrust while we are still trying to find our feet in the task of developing a satisfactory value theory. Nor is it always clear how to profit from the lessons such examples teach us when it comes to harder realistic ethical decision problems.

What I think this illustrates is that books like this may be getting very close to the limits of (relatively) informal reasoning, without the kind of formal and mathematical models with which most economists have become increasingly familiar during the course of the last fifty years. Such models can greatly facilitate the fitting together of rather complicated arguments, as well as ensuring logical consistency. Not that formal models are a complete answer by any means. One of the praiseworthy features of this book is how much more sensible and sophisticated Hardin appears to be in discussing economic policy issues than is much of the economics profession—especially some members of that profession who have also been associated with the University of Chicago. Formal models based on silly assumptions or silly ethical premises give silly formal conclusions, of course. But reasoning with the help of formal models can not only help avoid ambiguities and silly mistakes (although there are too many of these in this book): it can also aid in describing much more powerful versions of value theory, and enriched possibilities in game theory. Precision in thought is difficult without precision in language, and mathematical language is (usually) very precise.

VII. BEYOND PRESENT LIMITS

Future work in ethical theory cannot afford to ignore the many subtle arguments advanced in this book; I hope to see them refined using, where appropriate, the most recent developments in formal decision and game theory. Our ethical reasoning abilities still need to be greatly extended, and will be if we begin to take proper account of our human frailties. Readers of this journal are particularly well equipped to aid in this extension.

REFERENCES.

Bernheim, B. Douglas. 1984. “Rationalizable Strategic Behavior.” *Econometrica* 52: 1007–1028.

Bernheim, B. Douglas. 1986. “Axiomatic Characterizations of Rational Choice in Strategic Environments.” *Scandinavian Journal of Economics* 88: 473–488.

Blackorby, Charles and Donaldson, David. 1988. “Cash versus Kind, Self-Selection, and Efficient Transfers.” *American Economic Review* 78: 691–700.

Dasgupta, Partha S. 1982. “Utility, Information, and Rights.” In *Utilitarianism and Beyond* edited by Amartya K. Sen and Bernard Williams, pp. 199–218. Cambridge: Cambridge University Press.

Hammond, Peter J. 1990. “Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made.” Forthcoming in the Proceedings of the Sloan Conference on “Interpersonal Comparability of Welfare,” edited by Jon Elster and John Roemer. Cambridge: Cambridge University Press.

Hare, Richard M. 1981. *Moral Thinking: Its Levels, Method and Point*. Oxford: Clarendon Press.

Pearce, David. 1984. “Rationalizable Strategic Behavior and the Problem of Perfection.” *Econometrica* 52: 1029–1050.

Sen, Amartya K. 1970. *Collective Choice and Social Welfare.* San Francisco: Holden-Day.

Sen, Amartya K. 1977. “Rational Fools: A Critique of the Behavioural Foundations of Economic Theory.” *Philosophy and Public Affairs,* 6: 317–344.

Suppes, Patrick. 1966. “Some Formal Models of Grading Principles.” *Synthese*, 6: 284–306.