VI

CONTRACTUALISM AND UTILITARIANISM

T. M. SCANLON

Utilitarianism occupies a central place in the moral philosophy of our time. It is not the view which most people hold; certainly there are very few who would claim to be act-utilitarians. But for a much wider range of people it is the view towards which they find themselves pressed when they try to give a theoretical account of their moral beliefs. Within moral philosophy it represents a position one must struggle against if one wishes to avoid it. This is so in spite of the fact that the implications of act-utilitarianism are wildly at variance with firmly held moral convictions, while rule-utilitarianism, the most common alternative formulation, strikes most people as an unstable compromise.

The wide appeal of utilitarianism is due, I think, to philosophical considerations of a more or less sophisticated kind which pull us in a quite different direction than our first-order moral beliefs. In particular, utilitarianism derives much of its appeal from alleged difficulties about the foundations of rival views. What a successful alternative to utilitarianism must do, first and foremost, is to sap this source of strength by providing a clear account of the foundations of non-utilitarian moral reasoning. In what follows I will first describe the problem in more detail by setting out the questions which a philosophical account of the foundations of morality must answer. I will then put forward a version of contractualism which, I will argue, offers a better set of responses to these questions than that supplied by straightforward versions of utilitarianism. Finally I will explain why contractualism, as I understand it, does not lead back to some utilitarian formula as its normative outcome.

From Amartya Sen and Bernard Williams (eds.), *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982), 103–28. Reprinted by permission of Cambridge University Press and the author.

I am greatly indebted to Derek Parfit for patient criticism and enormously helpful discussion of many earlier versions of this paper. Thanks are due also to the many audiences who have heard parts of those versions delivered as lectures and kindly responded with helpful comments. In particular, I am indebted to Marshall Cohen, Ronald Dworkin, Owen Fiss, and Thomas Nagel for valuable criticism.

Contractualism has been proposed as the alternative to utilitarianism before, notably by John Rawls in *A Theory of Justice*.[1] Despite the wide discussion which this book has received, however, I think that the appeal of contractualism as a foundational view has been underrated. In particular, it has not been sufficiently appreciated that contractualism offers a particularly plausible account of moral motivation. The version of contractualism that I shall present differs from Rawls's in a number of respects. In particular, it makes no use, or only a different and more limited kind of use, of his notion of choice from behind a veil of ignorance. One result of this difference is to make the contrast between contractualism and utilitarianism stand out more clearly.

I

There is such a subject as moral philosophy for much the same reason that there is such a subject as the philosophy of mathematics. In moral judgements, as in mathematical ones, we have a set of putatively objective beliefs in which we are inclined to invest a certain degree of confidence and importance. Yet on reflection it is not at all obvious what, if anything, these judgements can be about, in virtue of which some can be said to be correct or defensible and others not. This question of subject-matter, or the grounds of truth, is the first philosophical question about both morality and mathematics. Second, in both morality and mathematics it seems to be possible to discover the truth simply by thinking or reasoning about it. Experience and observation may be helpful, but observation in the normal sense is not the standard means of discovery in either subject. So, given any positive answer to the first question—any specification of the subject-matter or ground of truth in mathematics or morality—we need some compatible epistemology explaining how it is possible to discover the facts about this subject-matter through something like the means we seem to use.

Given this similarity in the questions giving rise to moral philosophy and to the philosophy of mathematics, it is not surprising that the answers commonly given fall into similar general types. If we were to interview students in a freshman mathematics course many of them would, I think, declare themselves for some kind of conventionalism. They would hold that mathematics proceeds from definitions and principles that are either arbitrary or instrumentally justified, and that mathematical reasoning consists in perceiving what follows from these definitions and principles. A

[1] (Cambridge, Mass.: Harvard University Press, 1971.)

few others, perhaps, would be realists or platonists according to whom mathematical truths are a special kind of non-empirical fact that we can perceive through some form of intuition. Others might be naturalists who hold that mathematics, properly understood, is just the most abstract empirial science. Finally there are, though perhaps not in an average freshman course, those who hold that there are no mathematical facts in the world 'outside of us', but that the truths of mathematics are objective truths about the mental constructions of which we are capable. Kant held that pure mathematics was a realm of objective mind-dependent truths, and Brouwer's mathematical intuitionism is another theory of this type (with the important difference that it offers grounds for the warranted assertability of mathematical judgements rather than for their truth in the classical sense). All of these positions have natural correlates in moral philosophy. Intuitionism of the sort espoused by W. D. Ross is perhaps the closest analogue to mathematical platonism, and Kant's theory is the most familiar version of the thesis that morality is a sphere of objective, mind-dependent truths.

All of the views I have mentioned (with some qualification in the case of conventionalism) give positive (i.e. non-sceptical) answers to the first philosophical question about mathematics. Each identifies some objective, or at least intersubjective, ground of truth for mathematical judgements. Outright scepticism and subjective versions of mind-dependence (analogues of emotivism or prescriptivism) are less appealing as philosophies of mathematics than as moral philosophies. This is so in part simply because of the greater degree of intersubjective agreement in mathematical judgement. But it is also due to the difference in the further questions that philosophical accounts of the two fields must answer.

Neither mathematics nor morality can be taken to describe a realm of facts existing in isolation from the rest of reality. Each is supposed to be connected with other things. Mathematical judgements give rise to predictions about those realms to which mathematics is applied. This connection is something that a philosophical account of mathematical truth must explain, but the fact that we can observe and learn from the correctness of such predictions also gives support to our belief in objective mathematical truth. In the case of morality the main connection is, or is generally supposed to be, with the will. Given any candidate for the role of subject-matter of morality we must explain why anyone should care about it, and the need to answer this question of motivation has given strong support to subjectivist views.

But what must an adequate philosophical theory of morality say about moral motivation? It need not, I think, show that the moral truth gives anyone who knows it a reason to act which appeals to that person's present

desires or to the advancement of his or her interests. I find it entirely intelligible that moral requirement might correctly apply to a person even though that person had no reason of either of these kinds for complying with it. Whether moral requirements give those to whom they apply reasons for compliance of some third kind is a disputed question which I shall set aside. But what an adequate moral philosophy must do, I think, is to make clearer to us the nature of the reasons that morality does provide, at least to those who are concerned with it. A philosophical theory of morality must offer an account of these reasons that is, one the one hand, compatible with its account of moral truth and moral reasoning and, on the other, supported by a plausible analysis of moral experience. A satisfactory moral philosophy will not leave concern with morality as a simple special preference, like a fetish or a special taste, which some people just happen to have. It must make it understandable why moral reasons are ones that people can take seriously, and why they strike those who are moved by them as reasons of a special stringency and inescapability.

There is also a further question whether susceptibility to such reasons is compatible with a person's good or whether it is, as Nietzsche argued, a psychological disaster for the person who has it. If one is to defend morality one must show that it is not disastrous in this way, but I will not pursue this second motivational question here. I mention it only to distinguish it from the first question, which is my present concern.

The task of giving a philosophical explanation of the subject-matter of morality differs both from the task of analysing the meaning of moral terms and from that of finding the most coherent formulation of our first-order moral beliefs. A maximally coherent ordering of our first-order moral beliefs could provide us with a valuable kind of explanation: it would make clear how various, apparently disparate moral notions, precepts, and judgements are related to one another, thus indicating to what degree conflicts between them are fundamental and to what degree, on the other hand, they can be resolved or explained away. But philosophical inquiry into the subject-matter of morality takes a more external view. It seeks to explain what kind of truths moral truths are by describing them in relation to other things in the world and in relation to our particular concerns. An explanation of how we can come to know the truth about morality must be based on such an external explanation of the kind of things moral truths are rather than on a list of particular moral truths, even a maximally coherent list. This seems to be true as well about explanations of how moral beliefs can give one a reason to act.[2]

[2] Though here the ties between the nature of morality and its content are more important. It is not clear that an account of the nature of morality which left its content *entirely* open could be the basis for a plausible account of moral motivation.

Coherence among our first-order moral beliefs—what Rawls has called narrow reflective equilibrium[3]—seems unsatisfying[4] as an account of moral truth or as an account of the basis of justification in ethics just because, taken by itself, a maximally coherent account of our moral beliefs need not provide us with what I have called a philosophical explanation of the subject-matter of morality. However internally coherent our moral beliefs may be rendered, the nagging doubt may remain that there is nothing to them at all. They may be merely a set of socially inculcated reactions, mutually consistent perhaps but not judgements of a kind which can properly be said to be correct or incorrect. A philosophical theory of the nature of morality can contribute to our confidence in our first-order moral beliefs chiefly by allaying these natural doubts about the subject. In so far as it includes an account of moral epistemology, such a theory may guide us towards new forms of moral argument, but it need not do this. Moral argument of more or less the kind we have been familiar with may remain as the only form of justification in ethics. But whether or not it leads to revision in our modes of justification, what a good philosophical theory should do is to give us a clearer understanding of what the best forms of moral argument amount to and what kind of truth it is that they can be a way of arriving at. (Much the same can be said, I believe, about the contribution which philosophy of mathematics makes to our confidence in particular mathematical judgements and particular forms of mathematical reasoning.)

Like any thesis about morality, a philosophical account of the subject-matter of morality must have some connection with the meaning of moral terms: it must be plausible to claim that the subject-matter described is in fact what these terms refer to at least in much of their normal use. But the current meaning of moral terms is the product of many different moral beliefs held by past and present speakers of the language, and this meaning is surely compatible with a variety of moral views and with a variety of views about the nature of morality. After all, moral terms are used to express many different views of these kinds, and people who express these views are not using moral terms incorrectly, even though what some of them say must be mistaken. Like a first-order moral judgement, a philosophical characterization of the subject-matter of morality is a substantive claim about morality, albeit a claim of a different kind.

[3] See John Rawls, 'The Independence of Moral Theory', *Proceedings and Addresses of the American Philosophical Association*, 47 (1974–5), 8, and Norman Daniels, 'Wide Reflective Equilibrium and Theory Acceptance in Ethics', *Journal of Philosophy*, 76 (1979), 257–8. How closely the process of what I am calling philosophical explanation will coincide with the search for 'wide reflective equilibrium' as this is understood by Rawls and by Daniels is a further question which I cannot take up here.

[4] For expression of this dissatisfaction, see Peter Singer, 'Sidgwick and Reflective Equilibrium', *Monist*, 58 (1974), 490–517, and R. B. Brandt, *A Theory of the Good and the Right* (Oxford: Oxford University Press, 1979), 16–21.

While a philosophical characterization of morality makes a kind of claim that differs from a first-order moral judgement, this does not mean that a philosophical theory of morality will be neutral between competing normative doctrines. The adoption of a philosophical thesis about the nature of morality will almost always have some effect on the plausibility of particular moral claims, but philosophical theories of morality vary widely in the extent and directness of their normative implications. At one extreme is intuitionism, understood as the philosophical thesis that morality is concerned with certain non-natural properties. Rightness, for example, is held by Ross[5] to be the property of 'fittingness' or 'moral suitability'. Intuitionism holds that we can identify occurrences of these properties, and that we can recognize as self-evident certain general truths about them, but that they cannot be further analysed or explained in terms of other notions. So understood, intuitionism is in principle compatible with a wide variety of normative positions. One could, for example, be an intuitionistic utilitarian or an intuitionistic believer in moral rights, depending on the general truths about the property of moral rightness which one took to be self-evident.

The other extreme is represented by philosophical utilitarianism. The term 'utilitarianism' is generally used to refer to a family of specific normative doctrines—doctrines which might be held on the basis of a number of different philosophical theses about the nature of morality. In this sense of the term one might, for example, be a utilitarian on intuitionist or on contractualist grounds. But what I will call 'philosophical utilitarianism' is a particular philosophical thesis about the subject-matter of morality, namely the thesis that the only fundamental moral facts are facts about individual well-being.[6] I believe that this thesis has a great deal of plausibility for many people, and that, while some people are utilitarians for other reasons, it is the attractiveness of philosophical utilitarianism which accounts for the widespread influence of utilitarian principles.

It seems evident to people that there is such a thing as individuals' being made better or worse off. Such facts have an obvious motivational force; it is quite understandable that people should be moved by them in much the way that they are supposed to be moved by moral considerations. Further,

[5] W. D. Ross, *Foundations of Ethics* (Oxford: Oxford University Press, 1939), 52–4, 315.

[6] For purposes of this discussion I leave open the important questions of which individuals are to count and how 'well-being' is to be understood. Philosophical utilitarianism will retain the appeal I am concerned with under many different answers to these questions.

these facts are clearly relevant to morality as we now understand it. Claims about individual well-being are one class of valid starting-points for moral argument. But many people find it much harder to see how there could be any other, independent starting-points. Substantive moral requirements independent of individual well-being strike people as intuitionist in an objectionable sense. They would represent 'moral facts' of a kind it would be difficult to explain. There is no problem about recognizing it as a fact that a certain act is, say, an instance of lying or of promise-breaking. And a utilitarian can acknowledge that such facts as these often have (derivative) moral significance: they are morally significant because of their consequences for individual well-being. The problems, and the charge of 'intuitionism', arise when it is claimed that such acts are wrong in a sense that is not reducible to the fact that they decrease individual well-being. How could this independent property of moral wrongness be understood in a way that would give it the kind of importance and motivational force which moral considerations have been taken to have? If one accepts the idea that there are no moral properties having this kind of intrinsic significance, then philosophical utilitarianism may seem to be the only tenable account of morality. And once philosophical utilitarianism is accepted, some form of normative utilitarianism seems to be forced on us as the correct first-order moral theory. Utilitarianism thus has, for many people, something like the status which Hilbert's formalism and Brouwer's intuitionism have for their believers. It is a view which seems to be forced on us by the need to give a philosophically defensible account of the subject. But it leaves us with a hard choice: we can either abandon many of our previous first-order beliefs or try to salvage them by showing that they can be obtained as derived truths or explained away as useful and harmless fictions.

It may seem that the appeal of philosophical utilitarianism as I have described it is spurious, since this theory must amount either to a form of intuitionism (differing from others only in that it involves just one appeal to intuition) or else to definitional naturalism of a kind refuted by Moore and others long ago. But I do not think that the doctrine can be disposed of so easily. Philosophical utilitarianism is a philosophical thesis about the nature of morality. As such, it is on a par with intuitionism or with the form of contractualism which I will defend later in this paper. None of these theses need claim to be true as a matter of definition; if one of them is true it does not follow that a person who denies it is misusing the words 'right', 'wrong', and 'ought'. Nor are all these theses forms of intuitionism, if intuitionism is understood as the view that moral facts concern special non-natural properties, which we can apprehend by intuitive insight but which

do not need or admit of any further analysis. Both contractualism and philosophical utilitarianism are specifically incompatible with this claim. Like other philosophical theses about the nature of morality (including, I would say, intuitionism itself), contractualism and philosophical utilitarianism are to be appraised on the basis of their success in giving an account of moral belief, moral argument, and moral motivation that is compatible with our general beliefs about the world: our beliefs about what kinds of things there are in the world, what kinds of observation and reasoning we are capable of, and what kinds of reasons we have for action. A judgement as to which account of the nature of morality (or of mathematics) is most plausible in this general sense is just that: a judgement of overall plausibility. It is not usefully described as an insight into concepts or as a special intuitive insight of some other kind.

If philosophical utilitarianism is accepted then some form of utilitarianism appears to be forced upon us as a normative doctrine, but further argument is required to determine which form we should accept. If all that counts morally is the well-being of individuals, no one of whom is singled out as counting for more than the others, and if all that matters in the case of each individual is the degree to which his or her well-being is affected, then it would seem to follow that the basis of moral appraisal is the goal of maximizing the *sum*[7] of individual well-being. Whether this standard is to be applied to the criticism of individual actions, or to the selection of rules or policies, or to the inculcation of habits and dispositions to act is a further question, as is the question of how 'well-being' itself is to be understood. Thus the hypothesis that much of the appeal of utilitarianism as a normative doctrine derives from the attractiveness of philosophical utilitarianism explains how people can be convinced that some form of utilitarianism must be correct while yet being quite uncertain as to which form it is, whether it is 'direct' or 'act-' utilitarianism or some form of indirect 'rule-' or 'motive-' utilitarianism. What these views have in common, despite their differing normative consequences, is the identification of the same class of fundamental moral facts.

II

If what I have said about the appeal of utilitarianism is correct, then what a rival theory must do is to provide an alternative to philosophical utilitarianism as a conception of the subject-matter of morality. This is what the

[7] 'Average utilitarianism' is most plausibly arrived at through quite a different form of argument, one more akin to contractualism. I discuss one such argument in sect. IV below.

theory which I shall call contractualism seeks to do. Even if it succeeds in this, however, and is judged superior to philosophical utilitarianism as an account of the nature of morality, normative utilitarianism will not have been refuted. The possibility will remain that normative utilitarianism can be established on other grounds, for example as the normative outcome of contractualism itself. But one direct and, I think, influential argument for normative utilitarianism will have been set aside.

To give an example of what I mean by contractualism, a contractualist account of the nature of moral wrongness might be stated as follows. 'An act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behaviour which no one could reasonably reject as a basis for informed, unforced general agreement.' This is intended as a characterization of the kind of property which moral wrongness is. Like philosophical utilitarianism, it will have normative consequences, but it is not my present purpose to explore these in detail. As a contractualist account of one moral notion, what I have set out here is only an approximation, which may need to be modified considerably. Here I can offer a few remarks by way of clarification.

The idea of 'informed agreement' is meant to exclude agreement based on superstition or false belief about the consequences of actions, even if these beliefs are ones which it would be reasonable for the person in question to have. The intended force of the qualification 'reasonably', on the other hand, is to exclude rejections that would be unreasonable *given* the aim of finding principles which could be the basis of informed, unforced general agreement. Given this aim, it would be unreasonable, for example, to reject a principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others. I will have more to say about grounds for rejection later in the paper.

The requirement that the hypothetical agreement which is the subject of moral argument be unforced is meant not only to rule out coercion, but also to exclude being forced to accept an agreement by being in a weak bargaining position, for example because others are able to hold out longer and hence to insist on better terms. Moral argument abstracts from such considerations. The only relevant pressure for agreement comes from the desire to find and agree on principles which no one who had this desire could reasonably reject. According to contractualism, moral argument concerns the possibility of agreement among persons who are all moved by this desire, and moved by it to the same degree. But this counterfactual assumption characterizes only the agreement with which morality is concerned, not the world to which moral principles are to apply. Those who are concerned with morality look for principles for application to their

imperfect world which they could not reasonably reject, and which others in this world, who are not now moved by the desire for agreement, could not reasonably reject should they come to be so moved.[8]

The contractualist account of moral wrongness refers to principles 'which no one could reasonably reject' rather than to principles 'which everyone could reasonably accept' for the following reason.[9] Consider a principle under which some people will suffer severe hardships, and suppose that these hardships are avoidable. That is, there are alternative principles under which no one would have to bear comparable burdens. It might happen, however, that the people on whom these hardships fall are particularly self-sacrificing, and are willing to accept these burdens for the sake of what they see as the greater good of all. We would not say, I think, that it would be unreasonable of them to do this. On the other hand, it might not be unreasonable for them to refuse these burdens, and, hence, not unreasonable for someone to reject a principle requiring him to bear them. If this rejection would be reasonable, then the principle imposing these burdens is put in doubt, despite the fact that some particularly self-sacrificing people could (reasonably) accept it. Thus it is the reasonableness of rejecting a principle, rather than the reasonableness of accepting it, on which moral argument turns.

It seems likely that many non-equivalent sets of principles will pass the test of non-rejectability. This is suggested, for example, by the fact that there are many different ways of defining important duties, no one of which is more or less 'rejectable' than the others. There are, for example, many different systems of agreement-making and many different ways of assigning responsibility to care for others. It does not follow, however, that any action allowed by at least one of these sets of principles cannot be morally wrong according to contractualism. If it is important for us to have *some* duty of a given kind (some duty of fidelity to agreements, or some duty of mutual aid) of which there are many morally acceptable forms, then one of these forms needs to be established by convention. In a setting in which one of these forms *is* conventionally established, acts disallowed by it will be wrong in the sense of the definition given. For, given the need for such conventions, one thing that could not be generally agreed to would be a set of principles allowing one to disregard conventionally established (and morally acceptable) definitions of important duties. This dependence on convention introduces a degree of cultural relativity into contractualist morality. In addition, what a person can reasonably reject

[8] Here I am indebted to Gilbert Harman for comments which have helped me to clarify my statement of contractualism.

[9] A point I owe to Derek Parfit.

will depend on the aims and conditions that are important in his life, and these will also depend on the society in which he lives. The definition given above allows for variation of both of these kinds by making the wrongness of an action depend on the circumstances in which it is performed.

The partial statement of contractualism which I have given has the abstract character appropriate in an account of the subject-matter of morality. On its face, it involves no specific claim as to which principles could be agreed to or even whether there is a unique set of principles which could be the basis of agreement. One way, though not the only way, for a contractualist to arrive at substantive moral claims would be to give a technical definition of the relevant notion of agreement, e.g. by specifying the conditions under which agreement is to be reached, the parties to this agreement and the criteria of reasonableness to be employed. Different contractualists have done this in different ways. What must be claimed for such a definition is that (under the circumstances in which it is to apply) what it describes is indeed the kind of unforced, reasonable agreement at which moral argument aims. But contractualism can also be understood as an informal description of the subject-matter of morality on the basis of which ordinary forms of moral reasoning can be understood and appraised without proceeding via a technical notion of agreement.

Who is to be included in the general agreement to which contractualism refers? The scope of morality is a difficult question of substantive morality, but a philosophical theory of the nature of morality should provide some basis for answering it. What an adequate theory should do is to provide a framework within which what seem to be relevant arguments for and against particular interpretations of the moral boundary can be carried out. It is often thought that contractualism can provide no plausible basis for an answer to this question. Critics charge either that contractualism provides no answer at all, because it must begin with some set of contracting parties taken as given, or that contractualism suggests an answer which is obviously too restrictive, since a contract requires parties who are able to make and keep agreements and who are each able to offer the others some benefit in return for their co-operation. Neither of these objections applies to the version of contractualism that I defending. The general specification of the scope of morality which it implies seems to me to be this: morality applies to a being if the notion of justification to a being of that kind makes sense. What is required in order for this to be the case? Here I can only suggest some necessary conditions. The first is that the being have a good, that is, that there be a clear sense in which things can be said to go better or worse for that being. This gives partial sense to the idea of what it would be reasonable for a trustee to accept on the being's behalf. It would be

reasonable for a trustee to accept at least those things that are good, or not bad, for the being in question. Using this idea of trusteeship we can extend the notion of acceptance to apply to beings that are incapable of literally agreeing to anything. But this minimal notion of trusteeship is too weak to provide a basis for morality, according to contractualism. Contractualist morality relies on notions of what it would be reasonable to accept, or reasonable to reject, which are essentially comparative. Whether it would be unreasonable for me to reject a certain principle, given the aim of finding principles which no one with this aim could reasonably reject, depends not only on how much actions allowed by that principle might hurt me in absolute terms but also on how that potential loss compares with other potential losses to others under this principle and alternatives to it. Thus, in order for a being to stand in moral relations with us it is not enough that it have a good, it is also necessary that its good be sufficiently similar to our own to provide a basis for some system of comparability. Only on the basis of such a system can we give the proper kind of sense to the notion of what a trustee could reasonably reject on a being's behalf.

But the range of possible trusteeship is broader than that of morality. One could act as a trustee for a tomato plant, a forest, or an ant colony, and such entities are not included in morality. Perhaps this can be explained by appeal to the requirement of comparability: while these entities have a good, it is not comparable to our own in a way that provides a basis for moral argument. Beyond this, however, there is in these cases insufficient foothold for the notion of justification *to* a being. One further minimum requirement for this notion is that the being constitute a point of view: that is, that there be such a thing as what it is like to be that being, such a thing as what the world seems like to it. Without this, we do not stand in a relation to the being that makes even hypothetical justification *to it* appropriate.

On the basis of what I have said so far contractualism can explain why the capacity to feel pain should have seemed to many to count in favour of moral status: a being which has this capacity seems also to satisfy the three conditions I have just mentioned as necessary for the idea of justification to it to make sense. If a being can feel pain, then it constitutes a centre of consciousness to which justification can be addressed. Feeling pain is a clear way in which the being can be worse off; having its pain alleviated a way in which it can be benefited; and these are forms of weal and woe which seem directly comparable to our own.

It is not clear that the three conditions I have listed as necessary are also sufficient for the idea of justification to a being to make sense. Whether

they are, and, if they are not, what more may be required, are difficult and disputed questions. Some would restrict the moral sphere to those to whom justifications could in principle be communicated, or to those who can actually agree to something, or to those who have the capacity to understand moral argument. Contractualism as I have stated it does not settle these issues at once. All I claim is that it provides a basis for argument about them which is at least as plausible as that offered by rival accounts of the nature of morality. These proposed restrictions on the scope of morality are naturally understood as debatable claims about the conditions under which the relevant notion of justification makes sense, and the arguments commonly offered for and against them can also be plausibly understood on this basis.

Some other possible restrictions on the scope of morality are more evidently rejectable. Morality might be restricted to those who have the capacity to observe its constraints, or to those who are able to confer some reciprocal benefit on other participants. But it is extremely implausible to suppose that the beings excluded by these requirements fall entirely outside the protection of morality. Contractualism as I have formulated it[10] can explain why this is so: the absence of these capacities alone does nothing to undermine the possibility of justification to a being. What it may do in some cases, however, is to alter the justifications which are relevant. I suggest that whatever importance the capacities for deliberative control and reciprocal benefit may have is as factors altering the duties which beings have and the duties others have towards them, not as conditions whose absence suspends the moral framework altogether.

III

I have so far said little about the normative content of contractualism. For all I have said, the act-utilitarian formula might turn out to be a theorem of contractualism. I do not think that this is the case, but my main thesis is that whatever the normative implications of contractualism may be it still

[10] On this view (as contrasted with some others in which the notion of a contract is employed) what is fundamental to morality is the desire for reasonable agreement, not the pursuit of mutual advantage. See Sect. v below. It should be clear that this version of contractualism can account for the moral standing of future persons who will be better or worse off as a result of what we do now. It is less clear how it can deal with the problem presented by future people who would not have been born but for actions of ours which also made the conditions in which they live worse. Do such people have reason to reject principles allowing these actions to be performed? This difficult problem, which I cannot explore here, is raised by Derek Parfit in 'On Doing the Best for our Children', in M. Bayles (ed.), *Ethics and Population* (Cambridge, Mass.: Schenkman), 100–15.

has distinctive content as a philosophical thesis about the nature of morality. This content—the difference, for example, between being a utilitarian because the utilitarian formula is the basis of general agreement and being a utilitarian on other grounds—is shown most clearly in the answer that a contractualist gives to the first motivational question.

Philosophical utilitarianism is a plausible view partly because the facts which it identifies as fundamental to morality—facts about individual well-being—have obvious motivational force. Moral facts can motivate us, on this view, because of our sympathetic identification with the good of others. But as we move from philosophical utilitarianism to a specific utilitarian formula as the standard of right action, the form of motivation that utilitarianism appeals to becomes more abstract. If classical utilitarianism is the correct normative doctrine then the natural source of moral motivation will be a tendency to be moved by changes in aggregate well-being, however these may be composed. We must be moved in the same way by an aggregate gain of the same magnitude whether it is obtained by relieving the acute suffering of a few people or by bringing tiny benefits to a vast number, perhaps at the expense of moderate discomfort for a few. This is very different from sympathy of the familiar kind toward particular individuals, but a utilitarian may argue that this more abstract desire is what natural sympathy becomes when it is corrected by rational reflection. This desire has the same content as sympathy—it is a concern for the good of others—but it is not partial or selective in its choice of objects.

Leaving aside the psychological plausibility of this even-handed sympathy, how good a candidate is it for the role of moral motivation? Certainly sympathy of the usual kind is one of the many motives that can sometimes impel one to do the right thing. It may be the dominant motive, for example, when I run to the aid of a suffering child. But when I feel convinced by Peter Singer's article[11] on famine, and find myself crushed by the recognition of what seems a clear moral requirement, there is something else at work. In addition to the thought of how much good I could do for people in drought-stricken lands, I am overwhelmed by the further, seemingly distinct thought that it would be wrong for me to fail to aid them when I could do so at so little cost to myself. A utilitarian may respond that his account of moral motivation cannot be faulted for not capturing this aspect of moral experience, since it is just a reflection of our non-utilitarian moral upbringing. Moreover, it must be groundless. For what kind of fact could this supposed further fact of moral wrongness be, and how

[11] Peter Singer, 'Famine, Affluence, and Morality', *Philosophy and Public Affairs*, 1 (1972), 229–43.

could it give us a further, special reason for acting? The question for contractualism, then, is whether it can provide a satisfactory answer to this challenge.

According to contractualism, the source of motivation that is directly triggered by the belief that an action is wrong is the desire to be able to justify one's actions to others on grounds they could not reasonably[12] reject. I find this an extremely plausible account of moral motivation—a better account of at least my moral experience than the natural utilitarian alternative—and it seems to me to constitute a strong point for the contractualist view. We all might like to be in actual agreement with the people around us, but the desire which contractualism identifies as basic to morality does not lead us simply to conform to the standards accepted by others whatever these may be. The desire to be able to justify one's actions to others on grounds they could not reasonably reject will be satisfied when we know that there is adequate justification for our action even though others in fact refuse to accept it (perhaps because they have no interest in finding principles which we and others could not reasonably reject). Similarly, a person moved by this desire will not be satisfied by the fact that others accept a justification for his action if he regards this justification as spurious.

One rough test of whether you regard a justification as sufficient is whether you would accept that justification if you were in another person's position. This connection between the idea of 'changing places' and the motivation which underlies morality explains the frequent occurrence of 'Golden Rule' arguments within different systems of morality and in the teachings of various religions. But the thought experiment of changing places is only a rough guide; the fundamental question is what would it be unreasonable to reject as a basis for informed, unforced, general agreement. As Kant observed,[13] our different individual points of view, taken as they are, may in general be simply irreconcilable. 'Judgemental harmony' requires the construction of a genuinely interpersonal form of justification which is none the less something that each individual could agree to. From this interpersonal standpoint, a certain amount of how things look from another person's point of view, like a certain amount of how they look from my own, will be counted as bias.

I am not claiming that the desire to be able to justify one's actions to others on grounds they could not reasonably reject is universal or 'natural'.

[12] Reasonably, that is, given the desire to find principles which others similarly motivated could not reasonably reject.

[13] *Grundlegung zur Metaphysik der Sitten*, tr. H. J. Paton as *The Moral Law* (London: Hutchinson, 1948), sect. 2, n. 14.

'Moral education' seems to me plausibly understood as a process of cultivating this desire and shaping it, largely by learning what justifications others are in fact willing to accept, by finding which ones you yourself find acceptable as you confront them from a variety of perspectives, and by appraising your own and others' acceptance or rejection of these justifications in the light of greater experience.

In fact it seems to me that the desire to be able to justify one's actions (and institutions) on grounds one takes to be acceptable is quite strong in most people. People are willing to go to considerable lengths, involving quite heavy sacrifices, in order to avoid admitting the unjustifiability of their actions and institutions. The notorious insufficiency of moral motivation as a way of getting people to do the right thing is not due to simple weakness of the underlying motive, but rather to the fact that it is easily deflected by self-interest and self-deception.

It could reasonably be objected here that the source of motivation I have described is not tied exclusively to the contractualist notion of moral truth. The account of moral motivation which I have offered refers to the idea of a justification which it would be unreasonable to reject, and this idea is potentially broader than the contractualist notion of agreement. For let M be some non-contractualist account of moral truth. According to M, we may suppose, the wrongness of an action is simply a moral characteristic of that action in virtue of which it ought not to be done. An act which has this characteristic, according to M, has it quite independently of any tendency of informed persons to come to agreement about it. However, since informed persons are presumably in a position to recognize the wrongness of a type of action, it would seem to follow that if an action is wrong then such persons would agree that it is not to be performed. Similarly, if an act is not morally wrong, and there is adequate moral justification to perform it, then there will presumably be a moral justification for it which an informed person would be unreasonable to reject. Thus, even if M, and not contractualism, is the correct account of moral truth, the desire to be able to justify my actions to others on grounds they could not reasonably reject could still serve as a basis for moral motivation.

What this shows is that the appeal of contractualism, like that of utilitarianism, rests in part on a qualified scepticism. A non-contractualist theory of morality can make use of the source of motivation to which contractualism appeals. But a moral argument will trigger this source of motivation only in virtue of being a good justification for acting in a certain way, a justification which others would be unreasonable not to accept. So a non-contractualist theory must claim that there are moral properties which have justificatory force quite independent of their recognition in any

ideal agreement. These would represent what John Mackie has called instances of intrinsic 'to-be-doneness' and 'not-to-be-doneness'.[14] Part of contractualism's appeal rests on the view that, as Mackie puts it, it is puzzling how there could be such properties 'in the world'. By contrast, contractualism seeks to explain the justificatory status of moral properties, as well as their motivational force, in terms of the notion of reasonable agreement. In some cases the moral properties are themselves to be understood in terms of this notion. This is so, for example, in the case of the property of moral wrongness, considered above. But there are also right- and wrong-making properties which are themselves independent of the contractualist notion of agreement. I take the property of being an act of killing for the pleasure of doing so to be a wrong-making property of this kind. Such properties are wrong-making because it would be reasonable to reject any set of principles which permitted the acts they characterize. Thus, while there are morally relevant properties 'in the world' which are independent of the contractualist notion of agreement, these do not constitute instances of intrinsic 'to-be-doneness' and 'not-to-be-doneness': their moral relevance—their force in justifications as well as their link with motivation—is to be explained on contractualist grounds.

In particular, contractualism can account for the apparent moral significance of facts about individual well-being, which utilitarianism takes to be fundamental. Individual well-being will be morally significant, according to contractualism, not because it is intrinsically valuable or because promoting it is self-evidently a right-making characteristic, but simply because an individual could reasonably reject a form of argument that gave his well-being no weight. This claim of moral significance is, however, only approximate, since it is a further difficult question exactly how 'well-being' is to be understood and in what ways we are required to take account of the well-being of others in deciding what to do. It does not follow from this claim, for example, that a given desire will always and everywhere have the same weight in determining the rightness of an action that would promote its satisfaction, a weight proportional to its strength or 'intensity'. The right-making force of a person's desires is specified by what might be called a conception of morally legitimate interests. Such a conception is a product of moral argument: it is not given, as the notion of individual well-being may be, simply by the idea of what it is rational for an individual to desire. Not everything for which I have a rational desire will be something in which others need concede me to have a legitimate interest which they undertake to weigh in deciding what to do. The range of things which may

[14] J. L. Mackie, *Ethics: Inventing Right and Wrong* (Harmondsworth: Pelican, 1977), 42.

be objects of my rational desires is very wide indeed, and the range of claims which others could not reasonably refuse to recognize will almost certainly be narrower than this. There will be a tendency for interests to conform to rational desire—for those conditions making it rational to desire something also to establish a legitimate interest in it—but the two will not always coincide.

One effect of contractualism, then, is to break down the sharp distinction, which arguments for utilitarianism appeal to, between the status of individual well-being and that of other moral notions. A framework of moral argument is required to define our legitimate interests and to account for their moral force. This same contractualist framework can also account for the force of other moral notions such as rights, individual responsibility, and procedural fairness.

IV

It seems unlikely that act-utilitarianism will be a theorem of the version of contractualism which I have described. The positive moral significance of individual interests is a direct reflection of the contractualist requirement that actions be defensible to each person on grounds he could not reasonably reject. But it is a long step from here to the conclusion that each individual must agree to deliberate always from the point of view of maximum aggregate benefit and to accept justifications appealing to this consideration alone. It is quite possible that, according to contractualism, *some* moral questions may be properly settled by appeal to maximum aggregate well-being, even though this is not the sole or ultimate standard of justification.

What seems less improbable is that contractualism should turn out to coincide with some form of 'two-level' utilitarianism. I cannot fully assess this possibility here. Contractualism does share with these theories the important features that the defence of individual actions must proceed via a defence of principles that would allow those acts. But contractualism differs from *some* forms of two-level utilitarianism in an important way. The role of principles in contractualism is fundamental; they do not enter merely as devices for the promotion of acts that are right according to some other standard. Since it does not establish two potentially conflicting forms of moral reasoning, contractualism avoids the instability which often plagues rule-utilitarianism.

The fundamental question here, however, is whether the principles to which contractualism leads must be ones whose general adoption (either

ideally or under some more realistic conditions) would promote maximum aggregate well-being. It has seemed to many that this must be the case. To indicate why I do not agree I will consider one of the best known arguments for this conclusion and explain why I do not think it is successful. This will also provide an opportunity to examine the relation between the version of contractualism I have advocated here and the version set forth by Rawls.

The argument I will consider, which is familiar from the writings of Harsanyi[15] and others, proceeds via an interpretation of the contractualist notion of acceptance and leads to the principle of maximum average utility. To think of a principle as a candidate for unanimous agreement I must think of it not merely as acceptable to *me* (perhaps in virtue of my particular position, my tastes, etc.) but as acceptable[16] to others as well. To be relevant, my judgement that the principle is acceptable must be impartial. What does this mean? To judge impartially that a principle is acceptable is, one might say, to judge that it is one which you would have reason to accept no matter who you were. That is, and here is the interpretation, to judge that it is a principle which it would be rational to accept if you did not know which person's position you occupied and believed that you had an equal chance of being in any of these positions. ('Being in a person's position' is here understood to mean being in his objective circumstances and evaluating these from the perspective of his tastes and preferences.) But, it is claimed, the principle which it would be rational to prefer under these circumstances—the one which would offer the chooser greatest expected utility—would be that principle under which the average utility of the affected parties would be highest.

This argument might be questioned at a number of points, but what concerns me at present is the interpretation of impartiality. The argument can be broken down into three stages. The first of these is the idea that moral principles must be impartially acceptable. The second is the idea of choosing principles in ignorance of one's position (including one's tastes, preferences, etc.). The third is the idea of rational choice under the assumption that one has an equal chance of occupying anyone's position. Let me leave aside for the moment the move from stage two to stage three, and

[15] See John C. Harsanyi, 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economy*, 63 (1955), sect. IV. He is there discussing an argument which he presented earlier in Harsanyi, 'Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking', *Journal of Political Economy*, 61 (1953), 434–5.

[16] In discussing Harsanyi and Rawls I will generally follow them in speaking of the acceptability of principles rather than their unrejectability. The difference between these, pointed out above, is important only within the version of contractualism I am presenting; accordingly, I will speak of rejectability only when I am contrasting my own version with theirs.

concentrate on the first step, from stage one to stage two. There is a way of making something like this step which is, I think, quite valid, but it does not yield the conclusion needed by the argument. If I believe that a certain principle, *P*, could not reasonably be rejected as a basis for informed, unforced general agreement, then I must believe not only that it is something which it would be reasonable for me to accept but something which it would be reasonable for others to accept as well, in so far as we are all seeking a ground for general agreement. Accordingly, I must believe that I would have reason to accept *P* no matter which social position I were to occupy (though, for reasons mentioned above, I may not believe that I *would* agree to *P* if I were in some of these positions). Now it may be thought that no sense can be attached to the notion of choosing or agreeing to a principle in ignorance of one's social position, especially when this includes ignorance of one's tastes, preferences, etc. But there is at least a minimal sense that might be attached to this notion. If it would be reasonable for everyone to choose or agree to *P*, then my knowledge that I have reason to do so need not depend on my knowledge of my particular position, tastes, preferences, etc. So, in so far as it makes any sense at all to speak of choosing or agreeing to something in the absence of this knowledge, it could be said that I have reason to choose or agree to those things which everyone has reason to choose or agree to (assuming, again, the aim of finding principles on which all could agree). And indeed, this same reasoning can carry us through to a version of stage three. For if I judge *P* to be a principle which everyone has reason to agree to, then it could be said that I would have reason to agree to it if I thought that I had an equal chance of being anybody, or indeed, if I assign any other set of probabilities to being one or another of the people in question.

But it is clear that this is not the conclusion at which the original argument aimed. That conclusion concerned what it would be rational for a self-interested person to choose or agree to under the assumption of ignorance or equal probability of being anyone. The conclusion we have reached appeals to a different notion: the idea of what it would be unreasonable for people to reject given that they are seeking a basis for general agreement. The direction of explanation in the two arguments is quite different. The original argument sought to explain the notion of impartial acceptability of an ethical principle by appealing to the notion of rational self-interested choice under special conditions, a notion which appears to be a clearer one. My revised argument explains how *a* sense might be attached to the idea of choice or agreement in ignorance of one's position given some idea of what it would be unreasonable for someone to reject as a basis for general agreement. This indicates a problem for my

version of contractualism: it may be charged with failure to explain the central notion on which it relies. Here I would reply that my version of contractualism does not seek to explain this notion. It only tries to describe it clearly and to show how other features of morality can be understood in terms of it. In particular, it does not try to explain this notion by reducing it to the idea of what would maximize a person's self-interested expectations if he were choosing from a position of ignorance or under the assumption of equal probability of being anyone.

The initial plausibility of the move from stage one to stage two of the original argument rests on a subtle transition from one of these notions to the other. To believe that a principle is morally correct one must believe that it is one which all could reasonably agree to and none could reasonably reject. But my belief that this is the case may often be distorted by a tendency to take its advantage to me more seriously than its possible costs to others. For this reason, the idea of 'putting myself in another's place' is a useful corrective device. The same can be said for the thought experiment of asking what I could agree to in ignorance of my true position. But both of these thought experiments are devices for considering more accurately the question of what *everyone* could reasonably agree to or what no one could reasonably reject. That is, they involve the pattern of reasoning exhibited in my revised form of the three-stage argument, not that of the argument as originally given. The question, what would maximize the expectations of a single self-interested person choosing in ignorance of his true position, is a quite different question. This can be seen by considering the possibility that the distribution with the highest average utility, call it A, might involve extremely low utility levels for some people, levels much lower than the minimum anyone would enjoy under a more equal distribution.

Suppose that A is a principle which it would be rational for a self-interested chooser with an equal chance of being in anyone's position to select. Does it follow that no one could reasonably reject A? It seems evident that this does not follow.[17] Suppose that the situation of those who would fare worst under A, call them the Losers, is extremely bad, and that there is an alternative to A, call it E, under which no one's situation would be nearly as bad as this. Prima facie, the Losers would seem to have a reasonable ground for complaint against A. Their objection may be rebutted, by appeal to the sacrifices that would be imposed on some other

[17] The discussion which follows has much in common with the contrast between majority principles and unanimity principles drawn by Thomas Nagel in 'Equality', ch. 8 of *Mortal Questions* (Cambridge: Cambridge University Press, 1979). I am indebted to Nagel's discussion of this idea.

individual by the selection of E rather than A. But the mere fact that A yields higher average utility, which might be due to the fact that many people do very slightly better under A than under E while a very few do much worse, does not settle the matter.

Under contractualism, when we consider a principle our attention is naturally directed first to those who would do worst under it. This is because if anyone has reasonable grounds for objecting to the principle it is *likely* to be them. It does not follow, however, that contractualism always requires us to select the principle under which the expectations of the worse off are highest. The reasonableness of the Losers' objection to A is not established simply by the fact that they are worse off under A and no-one would be this badly off under E. The force of their complaint depends also on the fact that their position under A is, in absolute terms, very bad, and would be significantly better under E. This complaint must be weighed against those of individuals who would do worse under E. The question to be asked is, is it unreasonable for someone to refuse to put up with the Losers' situation under A in order that someone else should be able to enjoy the benefits which he would have to give up under E? As the supposed situation of the Loser under A becomes better, or his gain under E smaller in relation to the sacrifices required to produce it, his case is weakened.

One noteworthy feature of contractualist argument as I have presented it so far is that it is non-aggregative: what are compared are individual gains, losses, and levels of welfare. How aggregative considerations can enter into contractualist argument is a further question too large to be entered into here.

I have been criticizing an argument for average utilitarianism that is generally associated with Harsanyi, and my objections to this argument (leaving aside the last remarks about maximin) have an obvious similarity to objections raised by Rawls.[18] But the objections I have raised apply as well against some features of Rawls's own argument. Rawls accepts the first step of the argument I have described. That is, he believes that the correct principles of justice are those which 'rational persons concerned to advance their interests' would accept under the conditions defined by his original position, where they would be ignorant of their own particular talents, their conception of the good, and the social position (or generation) into which they were born. It is the second step of the argument which Rawls rejects, i.e. the claim that it would be rational for persons so

[18] For example, the intuitive argument against utilitarianism on p. 14 of Rawls, *A Theory of Justice*, and his repeated remark that we cannot expect some people to accept lower standards of life for the sake of the higher expectations of others.

situated to choose those principles which would offer them greatest expected utility under the assumption that they have an equal chance of being anyone in the society in question. I believe, however, that a mistake has already been made once the first step is taken.

This can be brought out by considering an ambiguity in the idea of acceptance by persons 'concerned to advance their interests'. On one reading, this is an essential ingredient in contractual argument; on another it is avoidable and, I think, mistaken. On the first reading, the interests in question are simply those of the members of society to whom the principles of justice are to apply (and by whom those principles must ultimately be accepted). The fact that they have interests which may conflict, and which they are concerned to advance, is what gives substance to questions of justice. On the second reading, the concern 'to advance their interests' that is in question is a concern of the parties to Rawls's original position, and it is this concern which determines, in the first instance,[19] what principles of justice they will adopt. Unanimous agreement among these parties, each motivated to do as well for himself as he can, is to be achieved by depriving them of any information that could give them reason to choose differently from one another. From behind the veil of ignorance, what offers the best prospects for one will offer the best prospects for all, since no one can tell what would benefit him in particular. Thus the choice of principles can be made, Rawls says, from the point of view of a single rational individual behind the veil of ignorance.

Whatever rules of rational choice this single individual, concerned to advance his own interests as best he can, is said to employ, this reduction of the problem to the case of a single person's self-interested choice should arouse our suspicion. As I indicated in criticizing Harsanyi, it is important to ask whether this single individual is held to accept a principle because he judges that it is one he could not reasonably reject whatever position he turns out to occupy, or whether, on the contrary, it is supposed to be acceptable to a person in any social position because it would be the rational choice for a single self-interested person behind the veil of ignorance. I have argued above that the argument for average utilitarianism involves a covert transition from the first pattern of reasoning to the second. Rawls's argument also appears to be of this second form; his defence of his two principles of justice relies, at least initially, on claims about what it would be rational for a person, concerned to advance his own

[19] 'Though they must then check to see that the principles they have chosen will be stable, not produce intolerable strains of commitment, and so on. As I argue below, these further considerations can be interpreted in a way that brings Rawls's theory closer to the version of contractualism presented here.

interests, to choose behind a veil of ignorance. I would claim, however, that the plausibility of Rawls's arguments favouring his two principles over the principle of average utility is preserved, and in some cases enhanced, when they are interpreted as instances of the first form of contractualist argument.

Some of these arguments are of an informal moral character. I have already mentioned his remark about the unacceptability of imposing lower expectations on some for the sake of the higher expectations of others. More specifically, he says of the parties to the original position that they are concerned 'to choose principles the consequences of which they are prepared to live with whatever generation they turn out to belong to'[20] or, presumably, whatever their social position turns out to be. This is a clear statement of the first form of contractualist argument. Somewhat later he remarks, in favour of the two principles, that they 'are those a person would choose for the design of a society in which his enemy is to assign him a place'.[21] Rawls goes on to dismiss this remark, saying that the parties 'should not reason from false premises',[22] but it is worth asking why it seemed a plausible thing to say in the first place. The reason, I take it, is this. In a contractualist argument of the first form, the object of which is to find principles acceptable to each person, assignment by a malevolent opponent is a thought experiment which has a heuristic role like that of a veil of ignorance: it is a way of testing whether one really does judge a principle to be acceptable from all points of view or whether, on the contrary, one is failing to take seriously its effect on people in social positions other than one's own.

But these are all informal remarks, and it is fair to suppose that Rawls's argument, like the argument for average utility, is intended to move from the informal contractualist idea of principles 'acceptable to all' to the idea of rational choice behind a veil of ignorance, an idea which is, he hopes, more precise and more capable of yielding definite results. Let me turn then to his more formal arguments for the choice of the difference principle by the parties to the original position. Rawls cites three features of the decision faced by parties to the original position which, he claims, make it rational for them to use the maximin rule and, therefore, to select his difference principle as a principle of justice. These are (1) the absence of any objective basis for estimating probabilities, (2) the fact that some principles could have consequences for them which 'they could hardly accept', while (3) it is possible for them (by following maximin) to ensure themselves of a minimum prospect, advances above which, in comparison,

[20] Rawls, *A Theory of Justice*, 137. [21] Ibid. 152. [22] Ibid. 153.

matter very little.[23] The first of these features is slightly puzzling, and I leave it aside. It seems clear, however, that the other considerations mentioned have at least as much force in an informal contractualist argument about what all could reasonably agree to as they do in determining the rational choice of a single person concerned to advance his interests. They express the strength of the objection that the 'losers' might have to a scheme that maximized average utility at their expense, as compared with the counter-objections that others might have to a more egalitarian arrangement.

In addition to this argument about rational choice, Rawls invokes among 'the main grounds for the two principles' other considerations which, as he says, use the concept of contract to a greater extent.[24] The parties to the original position, Rawls says, can agree to principles of justice only if they think that this agreement is one that they will actually be able to live up to. It is, he claims, more plausible to believe this of his two principles than of the principle of average utility, under which the sacrifices demanded ('the strains of commitment') could be much higher. A second, related claim is that the two principles of justice have greater psychological stability than the principle of average utility. It is more plausible to believe, Rawls claims, that in a society in which they were fulfilled people would continue to accept them and to be motivated to act in accordance with them. Continuing acceptance of the principle of average utility, on the other hand, would require an exceptional degree of identification with the good of the whole on the part of those from who sacrifices were demanded.

These remarks can be understood as claims about the 'stability' (in a quite practical sense) of a society founded on Rawls's two principles of justice. But they can also be seen as an attempt to show that a principle arrived at via the second form of contractualist reasoning will also satisfy the requirements of the first form, i.e. that it is something no one could reasonably reject. The question 'Is the acceptance of this principle an agreement you could actually live up to?' is, like the idea of assignment by one's worst enemy, a thought experiment through which we can use our own reactions to test our judgement that certain principles are ones that no one could reasonably reject. General principles of human psychology can also be invoked to this same end.

Rawls's final argument is that the adoption of his two principles gives public support to the self-respect of individual members of society, and 'give a stronger and more characteristic interpretation of Kant's idea'[25]

[23] Rawls, Ibid. 154. [24] Ibid., sect. 29, pp. 175 ff. [25] Ibid. 183.

that people must be treated as ends, not merely as means to the greater collective good. But, whatever difference there may be here between Rawls's two principles of justice and the principle of average utility, there is at least as sharp a contrast between the two patterns of contractualist reasoning distinguished above. The connection with self-respect, and with the Kantian formula, is preserved by the requirement that principles of justice be ones which no member of the society could reasonably reject. This connection is weakened when we shift to the idea of a choice which advances the interests of a single rational individual for whom the various individual lives in a society are just so many different possibilities. This is so whatever decision rule this rational chooser is said to employ. The argument from maximin seems to preserve this connection because it reproduces as a claim about rational choice what is, in slightly different terms, an appealing moral argument.

The 'choice situation' that is fundamental to contractualism as I have described it is obtained by beginning with 'mutually disinterested' individuals with full knowledge of their situations and adding to this (not, as is sometimes suggested, benevolence but) a desire on each of their parts to find principles which none could reasonably reject in so far as they too have this desire. Rawls several times considers such an idea in passing.[26] He rejects it in favour of his own idea of mutually disinterested choice from behind a veil of ignorance on the ground that only the latter enables us to reach definite results: 'if in choosing principles we required unanimity even where there is full information, only a few rather obvious cases could be decided'.[27] I believe that this supposed advantage is questionable. Perhaps this is because my expectations for moral argument are more modest than Rawls's. However, as I have argued, almost all of Rawls's own arguments have at least as much force when they are interpreted as arguments within the form of contractualism which I have been proposing. One possible exception is the argument from maximin. If the difference principle were taken to be generally applicable to decisions of public policy, then the second form of contractualist reasoning through which it is derived would have more far reaching implications than the looser form of argument by comparison of losses, which I have employed. But these wider applications of the principle are not always plausible, and I do not think that Rawls intends it to be applied so widely. His intention is that the difference principle should be applied only to major inequalities generated by the basic institutions of a society, and this limitation is a reflection of the

[26] e.g. ibid. 141, 148, although these passages may not clearly distinguish between this alternative and an assumption of benevolence.
[27] Ibid. 141.

special conditions under which he holds maximin to be the appropriate basis for rational choice: some choices have outcomes one could hardly accept, while gains above the minimum one can assure oneself matter very little, and so on. It follows, then, that in applying the difference principle—in identifying the limits of its applicability—we must fall back on the informal comparison of losses which is central to the form of contractualism I have described.

V

I have described this version of contractualism only in outline. Much more needs to be said to clarify its central notions and to work out its normative implications. I hope that I have said enough to indicate its appeal as a philosophical theory of morality and as an account of moral motivation. I have put forward contractualism as an alternative to utilitarianism, but the characteristic feature of the doctrine can be brought out by contrasting it with a somewhat different view.

It is sometimes said[28] that morality is a device for our mutual protection. According to contractualism, this view is partly true but in an important way incomplete. Our concern to protect our central interests will have an important effect on what we could reasonably agree to. It will thus have an important effect on the content of morality if contractualism is correct. To the degree that this morality is observed, these interests will gain from it. If we had no desire to be able to justify our actions to others on grounds they could reasonably accept, the hope of gaining this protection would give us reason to try to instil this desire in others, perhaps through mass hypnosis or conditioning, even if this also meant acquiring it ourselves. But given that we have this desire already, our concern with morality is less instrumental.

The contrast might be put as follows. On one view, concern with protection is fundamental, and general agreement becomes relevant as a means or a necessary condition for securing this protection. On the other, contractualist view, the desire for protection is an important factor determining the content of morality because it determines what can reasonably be agreed to. But the idea of general agreement does not arise as a means of securing protection. It is, in a more fundamental sense, what morality is about.

[28] In different ways by G. J. Warnock in *The Object of Morality* (London: Methuen, 1971), and by J. L. Mackie in *Ethics: Inventing Right and Wrong*. See also Richard Brandt's remarks on justification in ch. x of Brandt, *A Theory of the Good and the Right*.

VII

CAN THERE BE A RIGHT-BASED MORAL THEORY?

J. L. MACKIE

In the course of a discussion of Rawls's theory of justice, Ronald Dworkin suggests a 'tentative initial classification' of political theories into goal-based, right-based, and duty-based theories.[1] Though he describes this, too modestly, as superficial and trivial ideological sociology, it in fact raises interesting questions. In particular, does some such classification hold for moral as well as for political theories? We are familiar with goal-based or consequentialist moral views and with duty-based or deontological ones; but it is not easy to find right-based examples, and in discussions of consequentialism and deontology this third possibility is commonly ignored. Dworkin's own example of a right-based theory is Tom Paine's theory of revolution; another, recent, example might be Robert Nozick's theory of the minimal state.[2] But each of these is a political theory; the scope of each is restricted to the criticism of some political structures and policies and the support of others; neither is a fully developed general moral theory. If Rawls's view is, as Dworkin argues, fundamentally right-based, it may be the only member of this class. Moreover, it is only for Rawls's 'deep theory' that Dworkin can propose this identification: as explicitly formulated, Rawls's moral philosophy is not right-based. The lack of any convincing and decisive example leaves us free to ask the abstract question 'Could there be a right-based general moral theory, and, if there were one, what would it be like?'

It is obvious that most ordinary moral theories include theses about

Reprinted by permission of the University of Minnesota Press and Mrs J. Mackie, from Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein (eds.), *Studies in Ethical Theory*, Midwest Studies in Philosophy, 3 (Minneapolis: University of Minnesota Press, 1978), copyright © 1978 by the University of Minnesota.

[1] R. Dworkin, *Taking Rights Seriously* (London: Duckworth, 1977), ch. 6, 'Justice and Rights', esp. pp. 171–2. This chapter appeared first as an article, 'The Original Position', *University of Chicago Law Review*, 40 (1973), 500–33; repr. as ch. 2 in N. Daniels (ed.), *Reading Rawls* (Oxford: Oxford University Press, 1975).

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