A critique of utilitarianism

BERNARD WILLIAMS

If we possess our *why* of life we can put up with almost any *how*. – Man does not strive after happiness; only the Englishman does that.

Nietzsche, *The Twilight of the Idols*

1. Introductory

This essay is not designed as a reply to Smart's. It has been written after it, in knowledge of it, and from an opposed point of view, but it does not try to answer his arguments point for point, nor to cover just the same ground. Direct criticism of Smart's text is largely confined to parts of section 6, where I have tried to show that a certain ambiguity in Smart's defence of act-utilitarianism, as against other sorts, arises from a deep difficulty in the whole subject. I have not attempted, either, to give an account of all the important issues in the area, still less a critical survey of the major items in the literature; I have pursued those questions which seemed to me the most interesting and have deliberately left out a number of things which are often discussed. Like Smart, I have very largely treated utilitarianism as a system of personal morality rather than as a system of social or political decision, but I have tried to say something, very much in outline, about political aspects in section 7. The appearance of that subject at the end is not supposed to represent a judgement on its relative importance, but is due to two things: that I felt I had more to say about matters, such as those discussed in section 5, which bear most on the personal case; and that I think it important to come to the political area by a certain route, which involves the question "In whose hands does utilitarian decision lie?", and that route goes, I find, through the problems I consider in section 6 as arising for personal morality.

It is a merit of Smart's essay that it gives an account of utilitarianism which for the most part does not labour under

too many qualifications, and is only mildly apologetic. He thus stands in contrast to many modern writers whose utilitarianism is accommodated to a range of moral beliefs which many earlier utilitarians would probably have wanted to discard on the strength of utilitarianism. I agree with what in general is his stand (subject to the ambiguity I have mentioned, and which I discuss in section 6), that utilitarianism, properly understood and consistently carried through, is a *distinctive* way of looking at human action and morality. These distinctive characteristics he mostly seems to find agreeable, while to me some of them seem horrible. What is important, however (at least so far as these essays are concerned) is not whether he, or I, or the reader regard this or that as horrible, but what the implications, carefully considered, are of these principles for one's views of human nature and action, other people and society. Where I have offered examples, as particularly in section 3, the aim is not just to offer or elicit moral intuitions against which utilitarianism can be tested. Although in the end everyone has to reflect, in relation to questions like these, what he would be prepared to live with, the aim of the examples and their discussion is not just to ask a question about that and wait for the answer: rather, the aim is to lead into reflections which might show up in greater depth what would be involved in living with these ideas. The first question for philosophy is not "do you agree with utilitarianism's answer?" but "do you really accept utilitarianism's way of looking at the question?"

If utilitarianism is a distinctive moral outlook, that does not mean that there is just one way in which it is distinctive. If Smart's system is found by various critics, crass, or unjust, or muddled, or unrealistic, it may well be to different aspects of it that they are reacting, and I hope that my discussion will to some extent help to separate different strains of criticism of utilitarianism, and different features of utilitarian

systems to which they apply. There are three features in particular of Smart's system which may attract different kinds of criticism and which raise different kinds of issue. For these, reluctantly, I shall use some labels – reluctantly, because the use of technical labels in such matters can be a way of freezing the discussion, before one starts, into postures of antique controversy. But in this subject it is probably more misleading not to announce one's terminology, since many different technical terms, and different uses of the same terms to mark different distinctions, have been applied to it, and any term one uses will probably turn out to have been used by some other writer in a different sense. I shall be following some, at least, well-established practice in saying of Smart's system that it is *consequentialist*, and that its consequentialism is both *eudaimonistic* and *direct*.

Any kind of utilitarianism is by definition consequentialist, but 'consequentialism' is the broader term, and in my use (though not in everybody's use, and in particular, not in Smart's) utilitarianism is *one sort* of consequentialism – the sort (distinguished in the next paragraph) which is specially concerned with happiness. What is meant by 'consequentialism' turns out to be a harder question than at first appears, and I shall be concerned with it in section 2. It is also in my view an important question, since I think that some of the unacceptable features of utilitarianism, and some which I shall be particularly concerned with, are to be traced to its general character as a form of consequentialism. Very roughly speaking, consequentialism is the doctrine that the moral value of any action always lies in its consequences, and that it is by reference to their consequences that actions, and indeed such things as institutions, laws and practices, are to be justified if they can be justified at all.

To say, next, that the system is *eudaimonistic* is to say that what it regards as the desirable feature of actions is that they should increase or maximize people's *happiness*, as

distinguished from certain other goods at which, according to some consequentialists, it is independently worth aiming our actions. I shall not introduce any separate term to mark the view that the preferred value is *pleasure*, or again, *satisfaction*. Instead of talking about amounts of happiness, I shall sometimes use the economists' phrase, and speak of an increase or decrease in (people's) *utility*; and I shall in general assume, along with most modern writers in philosophy and economics, that in talking of happiness or utility one is talking about people's desires or preferences and their getting what they want or prefer, rather than about some sensation of pleasure or happiness. I say a little more about these matters in sections 2 and 3. The few remarks I have to make on the notorious problems of comparing and adding utilities, I have left to section 7; and for a good deal of the earlier discussion I have gone on as though this were not a problem. This is false, but the full force of its falsehood is felt, necessarily, at the level of social decision. It would be idle to pretend that in many more restricted connexions we had *no idea* what course would lead to greater happiness, and in earlier parts of the essay I have confined myself to difficulties which arise even when we can take that question as settled.

I shall rarely have to use the cumbrous term 'eudaimonistic' again, since I shall use the word 'utilitarianism' indeed to mean 'eudaimonistic consequentialism'. This is not Smart's practice, who uses the word 'utilitarianism' in the broader sense (and the phrase 'ideal utilitarianism' to refer to forms of consequentialism not exclusively concerned with happiness). His defence, indeed, ranges over these other sorts of consequentialism, but for much of the time he is concerned with what, in my narrower definition, is utilitarianism, that is to say, with consequentialism aimed at happiness. His various appeals to the principle of *benevolence* seem in particular to relate to that.

The term *direct* I use – putting it, again, very roughly – to mean that the consequential value which is the concern of morality is attached directly to particular actions, rather than to rules or practices under which decisions are taken without further reference to consequences; the latter sort of view is *indirect* consequentialism. The distinction, or one very like it, is often labelled, as it is by Smart, as a distinction between *act*-utilitarianism and *rule*-utilitarianism. I am sorry to have used a different terminology from Smart within the same covers, but in each case it proves simpler for my own purposes to do so; in the present matter, the term 'rule-utilitarianism' is less than useful, particularly because I am concerned with the indirect value of various sorts of things besides rules, such as dispositions. Like most other distinctions in this field, that between direct and indirect utilitarianism is easier to see at first glance than later, and it raises many complications. I consider some in section 6. I think, as Smart to some extent does, that forms of utilitarianism which help themselves too liberally to the resources of indirectness lose their utilitarian rationale and end up as vanishingly forms of utilitarianism at all. Whether that is so is not just a question of nomenclature or classification – such a question, in itself, would be of no interest at all. It is a question of the *point* of utilitarianism.[1]

This essay is concerned with utilitarianism, and in so far as it goes into consequentialism in general, this is only in order to suggest that some undesirable features of utilitarianism follow from its general consequentialist structure. Others follow more specifically from the nature of its concern with happiness. I shall say something about that,

[1] I have offered some brief arguments specifically related to that in *Morality: An Introduction to Ethics* (Harper and Row, New York, 1972; Penguin Books, Harmondsworth, 1973). Although there is some overlap between that treatment and the present essay, I have in general tried to develop rather different points.

and about the relations between direct and indirect forms of utilitarianism. I shall consider the uneasy relations of utilitarianism to certain other values which people either more or less optimistic than Smart might consider to have something seriously to do with human life. One value which has caused particular discomfort to utilitarianism is *justice*. I shall say a little about that in section 7, but I shall be more concerned with something rather different, *integrity*. I shall try to show something to which Smart's system indeed bears silent witness, that utilitarianism cannot hope to make sense, at any serious level, of integrity. It cannot do that for the very basic reason that it can make only the most superficial sense of human desire and action at all; and hence only very poor sense of what was supposed to be its own speciality, happiness.

2. The structure of consequentialism

No one can hold that everything, of whatever category, that has value, has it in virtue of its consequences. If that were so, one would just go on for ever, and there would be an obviously hopeless regress. That regress would be hopeless even if one takes the view, which is not an absurd view, that although men set themselves ends and work towards them, it is very often not really the supposed end, but the effort towards it on which they set value – that they travel, not really in order to arrive (for as soon as they have arrived they set out for somewhere else), but rather they choose somewhere to arrive, in order to travel. Even on that view, not everything would have consequential value; what would have non-consequential value would in fact be travelling, even though people had to think of travelling as having the consequential value, and something else – the destination – the non-consequential value.

If not everything that has value has it in virtue of consequences, then presumably there are some types of thing which have non-consequential value, and also some particular things that have such value because they are instances of those types. Let us say, using a traditional term, that anything that has that sort of value, has *intrinsic* value.[1] I take it to be the central idea of consequentialism that the only kind of thing that has intrinsic value is states of affairs, and that anything else that has value has it because it conduces to some intrinsically valuable state of affairs.

How much, however, does this say? Does it succeed in distinguishing consequentialism from anything else? The trouble is that the term 'state of affairs' seems altogether too permissive to exclude anything: may not the obtaining of absolutely anything be represented formally as a state of affairs? A Kantian view of morality, for instance, is usually thought to be opposed to consequentialism, if any is; at the very least, if someone were going to show that Kantianism collapsed into consequentialism, it should be the product of a long and unobvious argument, and not just happen at the drop of a definition. But on the present account it looks as though Kantianism can be made instantly into a kind of consequentialism – a kind which identifies the states of affairs that have intrinsic value (or at least intrinsic moral value) as those that consist of actions being performed for duty's sake.[2] We need something more to our specification if it is to be the specification of anything distinctly consequentialist.

The point of saying that consequentialism ascribes intrinsic value to states of affairs is rather to *contrast* states of

[1] The terminology of things 'being valuable', 'having intrinsic value', etc., is not meant to beg any questions in general value-theory. Non-cognitive theories, such as Smart's, should be able to recognize the distinctions made here.

[2] A point noted by Smart, p. 13.

affairs with other candidates for having such value: in particular, perhaps, actions. A distinctive mark of consequentialism might rather be this, that it regards the value of actions as always consequential (or, as we may more generally say, derivative), and not intrinsic. The value of actions would then lie in their causal properties, of producing valuable states of affairs; or if they did not derive their value in this simple way, they would derive it in some more roundabout way, as for instance by being expressive of some motive, or in accordance with some rule, whose operation in society conduced to desirable states of affairs. (The lengths to which such indirect derivations can be taken without wrecking the point of consequentialism is something we shall be considering later.)

To insist that what has intrinsic value are states of affairs and not actions seems to come near an important feature of consequentialism. Yet it may be that we have still not hit exactly what we want, and that the restriction is now too severe. Surely *some* actions, compatibly with consequentialism, might have intrinsic value? This is a question which has a special interest for utilitarianism, that is to say, the form of consequentialism concerned particularly with happiness. Traditionally utilitarians have tended to regard happiness or, again, pleasure, as experiences or sensations which were related to actions and activity as effect to cause; and, granted that view, utilitarianism will indeed see the value of all action as derivative, intrinsic value being reserved for the experiences of happiness. But that view of the relations between action and either pleasure or happiness is widely recognized to be inadequate. To say that a man finds certain actions or activity pleasant, or that they make him happy, or that he finds his happiness in them, is certainly not always to say that they induce certain sensations in him, and in the case of happiness, it is doubtful whether that is ever what is meant. Rather it means such things (among others) as that

he enjoys doing these things for their own sake. It would trivialize the discussion of utilitarianism to tie it by definition to inadequate conceptions of happiness or pleasure, and we must be able to recognize as versions of utilitarianism those which, as most modern versions do, take as central some notion such as *satisfaction*, and connect that criterially with such matters as the activities which a man will freely choose to engage in. But the activities which a man engages in for their own sake are activities in which he finds intrinsic value. So any specification of consequentialism which logically debars action or activity from having intrinsic value will be too restrictive even to admit the central case, utilitarianism, so soon as that takes on a more sophisticated and adequate conception of its basic value of happiness.

So far then, we seem to have one specification of consequentialism which is too generous to exclude anything, and another one which is too restrictive to admit even the central case. These difficulties arise from either admitting without question actions among desirable states of affairs, or blankly excluding all actions from the state of affairs category. This suggests that we shall do better by looking at the interrelations between states of affairs and actions.

It will be helpful, in doing this, to introduce the notion of the *right* action for an agent in given circumstances. I take it that in any form of direct consequentialism, and certainly in act-utilitarianism, the notion of the right action in given circumstances is a maximizing notion:[1] the right action is that which out of the actions available to the agent brings about or represents the highest degree of whatever it is the system in question regards as intrinsically valuable – in the central case, utilitarianism, this is of course happiness. In this argument, I shall confine myself to direct consequentialism, for which 'right action' is unqualifiedly a maximizing notion.

The notion of the right action as that which, of the possible

[1] Cf. Smart's definition, p. 45.

alternatives, maximizes the good (where this embraces, in unfavourable circumstances, minimizing the bad), is an objective notion in this sense, that it is perfectly possible for an agent to be ignorant or mistaken, and non-culpably ignorant or mistaken, about what is the right action in the circumstances. Thus the assessment by others of whether the agent did, in this sense, do the right thing, is not bounded by the agent's state of knowledge at the time, and the claim that he did the wrong thing is compatible with recognizing that he did as well as anyone in his state of knowledge could have done.[1] It might be suggested that, contrary to this, we have already imported the subjective conditions of action in speaking of the best of the actions *available to him*: if he is ignorant or misinformed, then the actions which might seem to us available to him were not in any real sense available. But this would be an exaggeration; the notion of availability imports some, but not all, kinds of subjective condition. Over and above the question of actions which, granted his situation and powers, were physically not available to him, we might perhaps add that a course of action was not really available to an agent if his historical, cultural or psychological situation was such that it could not possibly occur to him. But it is scarcely reasonable to extend the notion of unavailability to actions which merely did not occur to him; and surely absurd to extend it to actions which did occur to him, but where he was misinformed about their consequences.

If then an agent does the right thing, he does the best of the alternatives available to him (where that, again, embraces the least bad: we shall omit this rider from now on). Standardly, the action will be right in virtue of its causal properties, of maximally conducing to good states of affairs. Sometimes, however, the relation of the action to the good state of affairs may not be that of cause to effect – the good state

[1] In Smart's terminology, the 'rational thing': pp. 46–7.

of affairs may be constituted, or partly constituted, by the agent's doing that act (as when under utilitarianism he just enjoys doing it, and there is no project available to him more productive of happiness for him or anyone else).

Although this may be so under consequentialism, there seems to be an important difference between this situation and a situation of an action's being right for some non-consequentialist reason, as for instance under a Kantian morality. This difference might be brought out intuitively by saying that for the consequentialist, even a situation of this kind in which the action itself possesses intrinsic value is one in which the rightness of the act is derived from the goodness of a certain state of affairs – the act is right *because* the state of affairs which consists in its being done is better than any other state of affairs accessible to the agent; whereas for the non-consequentialist it is sometimes, at least, the other way round, and a state of affairs which is better than the alternatives is so because it consists of the right act being done. This intuitive description of the difference has something in it, but it needs to be made more precise.

We can take a step towards making it more precise, perhaps, in the following way. Suppose *S* is some particular concrete situation. Consider the statement, made about some particular agent

(1) In *S*, he did the right thing in doing *A*.

For consequentialists, (1) implies a statement of the form

(2) The state of affairs *P* is better than any other state of affairs accessible to him;

where a state of affairs being 'accessible' to an agent means that it is a state of affairs which is the consequence of, or is constituted by, his doing an act available to him (for that, see above); and *P* is a state of affairs accessible to him only in virtue of his doing *A*.[1]

[1] 'Only' here may seem a bit strong: but I take it that it is not an unreasonable demand on an account of his doing *the* right thing in *S* that his

Now in the exceptional case where it is just his doing *A* which carries the intrinsic value, we get for (2)

(3) The state of affairs which consists in his doing *A* is better than any other state of affairs accessible to him.

It was just the possibility of this sort of case which raised the difficulty of not being able to distinguish between a sophisticated consequentialism and non-consequentialism. The question thus is: if (3) is what we get for consequentialism in this sort of case, is it what a non-consequentialist would regard as implied by (1)? If so, we still cannot tell the difference between them. But the answer in fact seems to be 'no'.

There are two reasons for this. One reason is that a non-consequentialist, though he must inevitably be able to attach a sense to (1), does not have to be able to attach a sense to (3) at all, while the consequentialist, of course, attaches a sense to (1) only because he attaches a sense to (3). Although the non-consequentialist is concerned with right actions – such as the carrying out of promises – he may have no general way of comparing states of affairs from a moral point of view at all. Indeed, we shall see later and in greater depth than these schematic arguments allow, that the emphasis on the necessary comparability of situations is a peculiar feature of consequentialism in general, and of utilitarianism in particular.

A different kind of reason emerges if we suppose that the non-consequentialist does admit, in general, comparison between states of affairs. Thus, we might suppose that some non-consequentialist would consider it a better state of things in which more, rather than fewer, people kept their promises, and kept them for non-consequentialist reasons.

---

action is uniquely singled out from the alternatives. A further detail: one should strictly say, not that (1) implies a statement of the form (2), but that (1) implies *that there is* a true statement of that form.

Yet consistently with that he could accept, in a particular case, all of the following: that *X* would do the right thing only if he kept his promise; that keeping his promise would involve (or consist in) doing *A*; that several other people would, as a matter of fact, keep their promises (and for the right reasons) if and only if *X* did not do *A*. There are all sorts of situations in which this sort of thing would be true: thus it might be the case that an effect of *X*'s doing *A* would be to provide some inducement to these others which would lead them to break promises which otherwise they would have kept. Thus a non-consequentialist can hold both that it is a better state of affairs in which more people keep their promises, and that the right thing for *X* to do is something which brings it about that fewer promises are kept. Moreover, it is very obvious what view of things goes with holding that. It is one in which, even though from some abstract point of view one state of affairs is better than another, it does not follow that a given agent should regard it as his business to bring it about, even though it is open to him to do so. More than that, it might be that he could not properly regard it as his business. If the goodness of the world were to consist in people's fulfilling their obligations, it would by no means follow that one of my obligations was to bring it about that other people kept their obligations.

Of course, no sane person could really believe that the goodness of the world just consisted in people keeping their obligations. But that is just an example, to illustrate the point that under non-consequentialism (3) does not, as one might expect, follow from (1). Thus even allowing some actions to have intrinsic value, we can still distinguish consequentialism. A consequentialist view, then, is one in which a statement of the form (2) follows from a statement of the form (1). A non-consequentialist view is one in which this is not so – not even when the (2)-statement takes the special form of (3).

This is not at all to say that the alternative to consequentialism is that one has to accept that there are some actions which one should always do, or again some which one should never do, *whatever the consequences*: this is a much stronger position than any involved, as I have defined the issues, in the denial of consequentialism. All that is involved, on the present account, in the denial of consequentialism, is that with respect to some type of action, there are some situations in which that would be the right thing to do, even though the state of affairs produced by one's doing that would be worse than some other state of affairs accessible to one. The claim that there is a type of action which is right *whatever the consequences* can be put by saying that with respect to some type of action, assumed as being adequately specified, then *whatever* the situation may (otherwise) be, that will be the right thing to do, *whatever* other state of affairs might be accessible to one, however much better it might be than the state of affairs produced by one's doing this action.

If that somewhat Moorean formulation has not hopelessly concealed the point, it will be seen that this second position – the *whatever the consequences* position – is very much stronger than the first, the mere rejection of consequentialism. It is perfectly consistent, and it might be thought a mark of sense, to believe, while not being a consequentialist, that there was no type of action which satisfied this second condition: that if an adequate (and non-question-begging) specification of a type of action has been given in advance, it is always possible to think of some situation in which the consequences of doing the action so specified would be so awful that it would be right to do something else.

Of course, one might think that there just *were* some types of action which satisfied this condition; though it seems to me obscure how one could have much faith in a list of such actions unless one supposed that it had supernatural warrant.

Alternatively, one might think that while logically there was a difference between the two positions, in social and psychological fact they came to much the same thing, since so soon (it might be claimed) as people give up thinking in terms of certain things being right or wrong whatever the consequences, they turn to thinking in purely consequential terms. This might be offered as a very general proposition about human thought, or (more plausibly) as a sociological proposition about certain situations of social change, in which utilitarianism (in particular) looks the only coherent alternative to a dilapidated set of values. At the level of language, it is worth noting that the use of the word '*absolute*' mirrors, and perhaps also assists, this association: the claim that no type of action is 'absolutely right' - leaving aside the sense in which it means that the rightness of anything depends on the value-system of a society (the confused doctrine of relativism) - can mean either that no type of action is right-whatever-its-consequences, or, alternatively, that 'it all depends on the consequences', that is, in each case the decision whether an action is right is determined by its consequences.

A particular sort of psychological connexion - or in an old-fashioned use of the term, a 'moral' connexion - between the two positions might be found in this. If people do not regard certain things as 'absolutely out', then they are prepared to start thinking about extreme situations in which what would otherwise be out might, exceptionally, be justified. They will, if they are to get clear about what they believe, be prepared to compare different extreme situations and ask what action would be justified in them. But once they have got used to that, their inhibitions about thinking of everything in consequential terms disappear: the difference between the extreme situations and the less extreme, presents itself no longer as a difference between the exceptional and the usual, but between the greater and the less -

and the consequential thoughts one was prepared to deploy in the greater it may seem quite irrational not to deploy in the less. *A fortiori*, someone might say: but he would have already had to complete this process to see it as a case of *a fortiori*.

One could regard this process of adaptation to consequentialism, moreover, not merely as a blank piece of psychological association, but as concealing a more elaborate structure of thought. One might have the idea that the *unthinkable* was itself a moral category; and in more than one way. It could be a feature of a man's moral outlook that he regarded certain courses of action as unthinkable, in the sense that he would not entertain the idea of doing them: and the witness to that might, in many cases, be that they simply would not come into his head. Entertaining certain alternatives, regarding them indeed as *alternatives*, is itself something that he regards as dishonourable or morally absurd. But, further, he might equally find it unacceptable to consider what to do in certain conceivable situations. Logically, or indeed empirically conceivable they may be, but they are not to him morally conceivable, meaning by that that their occurrence as situations presenting him with a choice would represent not a special problem in his moral world, but something that lay beyond its limits. For him, there are certain situations so monstrous that the idea that the processes of moral rationality could yield an answer in them is insane: they are situations which so transcend in enormity the human business of moral deliberation that from a moral point of view it cannot matter any more what happens. Equally, for him, to spend time thinking what one would decide if one were in such a situation is also insane, if not merely frivolous.

For such a man, and indeed for anyone who is prepared to take him seriously, the demand, in Herman Kahn's words, to *think the unthinkable* is not an unquestionable demand of rationality, set against a cowardly or inert refusal to follow

out one's moral thoughts. Rationality he sees as a demand not merely on him, but on the situations in, and about, which he has to think; unless the environment reveals minimum sanity, it is insanity to carry the decorum of sanity into it. Consequentialist rationality, however, and in particular utilitarian rationality, has no such limitations: making the best of a bad job is one of its maxims, and it will have something to say even on the difference between massacring seven million, and massacring seven million and one.

There are other important questions about the idea of the morally unthinkable, which we cannot pursue here. Here we have been concerned with the role it might play in someone's connecting, by more than a mistake, the idea that there was nothing which was right whatever the consequences, and the different idea that everything depends on consequences. While someone might, in this way or another, move from one of those ideas to the other, it is very important that the two ideas are different: especially important in a world where we have lost traditional reasons for resisting the first idea, but have more than enough reasons for fearing the second.

3. Negative responsibility: and two examples

Although I have defined a state of affairs being *accessible* to an agent in terms of the actions which are *available* to him,[1] nevertheless it is the former notion which is really more important for consequentialism. Consequentialism is basically indifferent to whether a state of affairs consists in what I do, or is produced by what I do, where that notion is itself wide enough to include, for instance, situations in which other people do things which I have made them do, or allowed them to do, or encouraged them to do, or given them a chance to do. All that consequentialism is interested in is

[1] See last section, p. 87.

the idea of these doings being *consequences* of what I do, and that is a relation broad enough to include the relations just mentioned, and many others.

Just what the relation is, is a different question, and at least as obscure as the nature of its relative, cause and effect. It is not a question I shall try to pursue; I will rely on cases where I suppose that any consequentialist would be bound to regard the situations in question as consequences of what the agent does. There are cases where the supposed consequences stand in a rather remote relation to the action, which are sometimes difficult to assess from a practical point of view, but which raise no very interesting question for the present enquiry. The more interesting points about consequentialism lie rather elsewhere. There are certain situations in which the causation of the situation, the relation it has to what I do, is in no way remote or problematic in itself, and entirely justifies the claim that the situation is a consequence of what I do: for instance, it is quite clear, or reasonably clear, that if I do a certain thing, this situation will come about, and if I do not, it will not. So from a consequentialist point of view it goes into the calculation of consequences along with any other state of affairs accessible to me. Yet from some, at least, non-consequentialist points of view, there is a vital difference between some such situations and others: namely, that in some a vital link in the production of the eventual outcome is provided by *someone else's* doing something. But for consequentialism, all causal connexions are on the same level, and it makes no difference, so far as that goes, whether the causation of a given state of affairs lies through another agent, or not.

Correspondingly, there is no relevant difference which consists *just* in one state of affairs being brought about by me, without intervention of other agents, and another being brought about through the intervention of other agents; although some genuinely causal differences involving a

difference of value may correspond to that (as when, for instance, the other agents derive pleasure or pain from the transaction), that kind of difference will already be included in the specification of the state of affairs to be produced. Granted that the states of affairs have been adequately described in causally and evaluatively relevant terms, it makes no further comprehensible difference who produces them. It is because consequentialism attaches value ultimately to states of affairs, and its concern is with what states of affairs the world contains, that it essentially involves the notion of *negative responsibility*: that if I am ever responsible for anything, then I must be just as much responsible for things that I allow or fail to prevent, as I am for things that I myself, in the more everyday restricted sense, bring about.[1] Those things also must enter my deliberations, as a responsible moral agent, on the same footing. What matters is what states of affairs the world contains, and so what matters with respect to a given action is what comes about if it is done, and what comes about if it is not done, and those are questions not intrinsically affected by the nature of the causal linkage, in particular by whether the outcome is partly produced by other agents.

The strong doctrine of negative responsibility flows directly from consequentialism's assignment of ultimate value to states of affairs. Looked at from another point of view, it can be seen also as a special application of something that is favoured in many moral outlooks not themselves consequentialist – something which, indeed, some thinkers have been disposed to regard as the essence of morality

[1] This is a fairly modest sense of 'responsibility', introduced merely by one's ability to reflect on, and decide, what one ought to do. This presumably escapes Smart's ban (p. 54) on the notion of 'the responsibility' as 'a piece of metaphysical nonsense' – his remarks seem to be concerned solely with situations of inter-personal blame. For the limitations of that, see below, section 6 (pp. 123 ff.).

itself: a principle of impartiality. Such a principle will claim that there can be no relevant difference from a moral point of view which consists just in the fact, not further explicable in general terms, that benefits or harms accrue to one person rather than to another – 'it's me' can never in itself be a morally comprehensible reason.[1] This principle, familiar with regard to the reception of harms and benefits, we can see consequentialism as extending to their production: from the moral point of view, there is no comprehensible difference which consists just in my bringing about a certain outcome rather than someone else's producing it. That the doctrine of negative responsibility represents in this way the extreme of impartiality, and abstracts from the identity of the agent, leaving just a locus of causal intervention in the world – that fact is not merely a surface paradox. It helps to explain why consequentialism can seem to some to express a more serious attitude than non-consequentialist views, why part of its appeal is to a certain kind of high-mindedness. Indeed, that is part of what is wrong with it.

For a lot of the time so far we have been operating at an exceedingly abstract level. This has been necessary in order to get clearer in general terms about the differences between consequentialist and other outlooks, an aim which is important if we want to know what features of them lead to what results for our thought. Now, however, let us look more concretely at two examples, to see what utilitarianism might say about them, what we might say about utilitarianism and, most importantly of all, what would be implied by certain ways of thinking about the situations. The examples are inevitably schematized, and they are open to the objection that they beg as many questions as they illuminate. There are two ways in particular in which examples in

[1] There is a tendency in some writers to suggest that it is not a comprehensible reason at all. But this, I suspect, is due to the overwhelming importance those writers ascribe to the moral point of view.

moral philosophy tend to beg important questions. One is that, as presented, they arbitrarily cut off and restrict the range of alternative courses of action – this objection might particularly be made against the first of my two examples. The second is that they inevitably present one with the situation as a going concern, and cut off questions about how the agent got into it, and correspondingly about moral considerations which might flow from that: this objection might perhaps specially arise with regard to the second of my two situations. These difficulties, however, just have to be accepted, and if anyone finds these examples cripplingly defective in this sort of respect, then he must in his own thought rework them in richer and less question-begging form. If he feels that no presentation of any imagined situation can ever be other than misleading in morality, and that there can never be any substitute for the concrete experienced complexity of actual moral situations, then this discussion, with him, must certainly grind to a halt: but then one may legitimately wonder whether every discussion with him about conduct will not grind to a halt, including any discussion about the actual situations, since discussion about how one would think and feel about situations somewhat different from the actual (that is to say, situations to that extent imaginary) plays an important role in discussion of the actual.

(1) George, who has just taken his Ph.D. in chemistry, finds it extremely difficult to get a job. He is not very robust in health, which cuts down the number of jobs he might be able to do satisfactorily. His wife has to go out to work to keep them, which itself causes a great deal of strain, since they have small children and there are severe problems about looking after them. The results of all this, especially on the children, are damaging. An older chemist, who knows about this situation, says that he can get George a decently paid job in a certain laboratory, which pursues research into

chemical and biological warfare. George says that he cannot accept this, since he is opposed to chemical and biological warfare. The older man replies that he is not too keen on it himself, come to that, but after all George's refusal is not going to make the job or the laboratory go away; what is more, he happens to know that if George refuses the job, it will certainly go to a contemporary of George's who is not inhibited by any such scruples and is likely if appointed to push along the research with greater zeal than George would. Indeed, it is not merely concern for George and his family, but (to speak frankly and in confidence) some alarm about this other man's excess of zeal, which has led the older man to offer to use his influence to get George the job . . . George's wife, to whom he is deeply attached, has views (the details of which need not concern us) from which it follows that at least there is nothing particularly wrong with research into CBW. What should he do?

(2) Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition, explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protestors of the advantages of not protesting. However, since Jim is an honoured visitor from another land, the captain is happy to offer him a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all. Jim, with some desperate recollec-

tion of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro and the rest of the soldiers to threat, but it is quite clear from the set-up that nothing of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do?

To these dilemmas, it seems to me that utilitarianism replies, in the first case, that George should accept the job, and in the second, that Jim should kill the Indian. Not only does utilitarianism give these answers but, if the situations are essentially as described and there are no further special factors, it regards them, it seems to me, as *obviously* the right answers. But many of us would certainly wonder whether, in (1), that could possibly be the right answer at all; and in the case of (2), even one who came to think that perhaps that was the answer, might well wonder whether it was obviously the answer. Nor is it just a question of the rightness or obviousness of these answers. It is also a question of what sort of considerations come into finding the answer. A feature of utilitarianism is that it cuts out a kind of consideration which for some others makes a difference to what they feel about such cases: a consideration involving the idea, as we might first and very simply put it, that each of us is specially responsible for what *he* does, rather than for what other people do. This is an idea closely connected with the value of integrity. It is often suspected that utilitarianism, at least in its direct forms, makes integrity as a value more or less unintelligible. I shall try to show that this suspicion is correct. Of course, even if that is correct, it would not necessarily follow that we should reject utilitarianism; perhaps, as utilitarians sometimes suggest, we should just forget about integrity, in favour of such things as a concern for the general good. However, if I am right,

we cannot merely do that, since the reason why utilitarianism cannot understand integrity is that it cannot coherently describe the relations between a man's projects and his actions.

4. Two kinds of remoter effect

A lot of what we have to say about this question will be about the relations between my projects and other people's projects. But before we get on to that, we should first ask whether we are assuming too hastily what the utilitarian answers to the dilemmas will be. In terms of more direct effects of the possible decisions, there does not indeed seem much doubt about the answer in either case; but it might be said that in terms of more remote or less evident effects counterweights might be found to enter the utilitarian scales. Thus the effect on George of a decision to take the job might be invoked, or its effect on others who might know of his decision. The possibility of there being more beneficent labours in the future from which he might be barred or disqualified, might be mentioned; and so forth. Such effects – in particular, possible effects on the agent's character, and effects on the public at large – are often invoked by utilitarian writers dealing with problems about lying or promise-breaking, and some similar considerations might be invoked here.

There is one very general remark that is worth making about arguments of this sort. The certainty that attaches to these hypotheses about possible effects is usually pretty low; in some cases, indeed, the hypothesis invoked is so implausible that it would scarcely pass if it were not being used to deliver the respectable moral answer, as in the standard fantasy that one of the effects of one's telling a particular lie is to weaken the disposition of the world at large to tell the truth. The demands on the certainty or probability of these beliefs as beliefs about particular actions are much milder

than they would be on beliefs favouring the unconventional course. It may be said that this is as it should be, since the presumption must be in favour of the conventional course: but that scarcely seems a *utilitarian* answer, unless utilitarianism has already taken off in the direction of not applying the consequences to the particular act at all.

Leaving aside that very general point, I want to consider now two types of effect that are often invoked by utilitarians, and which might be invoked in connexion with these imaginary cases. The attitude or tone involved in invoking these effects may sometimes seem peculiar; but that sort of peculiarity soon becomes familiar in utilitarian discussions, and indeed it can be something of an achievement to retain a sense of it.

First, there is the psychological effect on the agent. Our descriptions of these situations have not so far taken account of how George or Jim will be after they have taken the one course or the other; and it might be said that if they take the course which seemed at first the utilitarian one, the effects on them will be in fact bad enough and extensive enough to cancel out the initial utilitarian advantages of that course. Now there is one version of this effect in which, for a utilitarian, some confusion must be involved, namely that in which the agent feels bad, his subsequent conduct and relations are crippled and so on, *because he thinks that he has done the wrong thing* – for if the balance of outcomes was as it appeared to be *before* invoking this effect, then he has not (from the utilitarian point of view) done the wrong thing. So that version of the effect, for a rational and utilitarian agent, could not possibly make any difference to the assessment of right and wrong. However, perhaps he is not a thoroughly rational agent, and is disposed to have bad feelings, whichever he decided to do. Now such feelings, which are from a strictly utilitarian point of view irrational – nothing, a utilitarian can point out, is advanced by having

them – cannot, consistently, have any great weight in a utilitarian calculation. I shall consider in a moment an argument to suggest that they should have no weight at all in it. But short of that, the utilitarian could reasonably say that such feelings should not be encouraged, even if we accept their existence, and that to give them a lot of weight is to encourage them. Or, at the very best, even if they are straightforwardly and without any discount to be put into the calculation, their weight must be small: they are after all (and at best) one man's feelings.

That consideration might seem to have particular force in Jim's case. In George's case, his feelings represent a larger proportion of what is to be weighed, and are more commensurate in character with other items in the calculation. In Jim's case, however, his feelings might seem to be of very little weight compared with other things that are at stake. There is a powerful and recognizable appeal that can be made on this point: as that a refusal by Jim to do what he has been invited to do would be a kind of self-indulgent squeamishness. That is an appeal which can be made by other than utilitarians – indeed, there are some uses of it which cannot be consistently made by utilitarians, as when it essentially involves the idea that there is something dishonourable about such self-indulgence. But in some versions it is a familiar, and it must be said a powerful, weapon of utilitarianism. One must be clear, though, about what it can and cannot accomplish. The most it can do, so far as I can see, is to invite one to consider how seriously, and for what reasons, one feels that what one is invited to do is (in these circumstances) wrong, and in particular, to consider that question from the utilitarian point of view. When the agent is not seeing the situation from a utilitarian point of view, the appeal cannot force him to do so; and if he does come round to seeing it from a utilitarian point of view, there is virtually nothing left for the appeal to do. If he does not see

it from a utilitarian point of view, he will not see his resistance to the invitation, and the unpleasant feelings he associates with accepting it, *just* as disagreeable experiences of his; they figure rather as emotional expressions of a thought that to accept would be wrong. He may be asked, as by the appeal, to consider whether he is right, and indeed whether he is fully serious, in thinking that. But the assertion of the appeal, that he is being self-indulgently squeamish, will not itself answer that question, or even help to answer it, since it essentially tells him to regard his feelings just as unpleasant experiences of his, and he cannot, by doing that, answer the question they pose when they are precisely not so regarded, but are regarded as indications[1] of what he thinks is right and wrong. If he does come round fully to the utilitarian point of view then of course he will regard these feelings just as unpleasant experiences of his. And once Jim - at least - has come to see them in that light, there is nothing left for the appeal to do, since *of course* his feelings, so regarded, are of virtually no weight at all in relation to the other things at stake. The 'squeamishness' appeal is not an argument which adds in a hitherto neglected consideration. Rather, it is an invitation to consider the situation, and one's own feelings, from a utilitarian point of view.

The reason why the squeamishness appeal can be very unsettling, and one can be unnerved by the suggestion of self-indulgence in going against utilitarian considerations, is not that we are utilitarians who are uncertain what utilitarian value to attach to our moral feelings, but that we are partially at least not utilitarians, and cannot regard our moral feelings merely as objects of utilitarian value. Because our moral relation to the world is partly given by such feelings, and by a sense of what we can or cannot 'live with', to come

[1] On the non-cognitivist meta-ethic in terms of which Smart presents his utilitarianism, the term 'indications' here would represent an understatement.

to regard those feelings from a purely utilitarian point of view, that is to say, as happenings outside one's moral self, is to lose a sense of one's moral identity; to lose, in the most literal way, one's integrity. At this point utilitarianism alienates one from one's moral feelings; we shall see a little later how, more basically, it alienates one from one's actions as well.

If, then, one is really going to regard one's feelings from a strictly utilitarian point of view, Jim should give very little weight at all to his; it seems almost indecent, in fact, once one has taken that point of view, to suppose that he should give any at all. In George's case one might feel that things were slightly different. It is interesting, though, that one reason why one might think that – namely that one person principally affected is his wife – is very dubiously available to a utilitarian. George's wife has some reason to be interested in George's integrity and his sense of it; the Indians, quite properly, have no interest in Jim's. But it is not at all clear how utilitarianism would describe that difference.

There is an argument, and a strong one, that a strict utilitarian should give not merely small extra weight, in calculations of right and wrong, to feelings of this kind, but that he should give absolutely no weight to them at all. This is based on the point, which we have already seen, that if a course of action is, before taking these sorts of feelings into account, utilitarianly preferable, then bad feelings about that kind of action will be from a utilitarian point of view irrational. Now it might be thought that even if that is so, it would not mean that in a utilitarian calculation such feelings should not be taken into account; it is after all a well-known boast of utilitarianism that it is a realistic outlook which seeks the best in the world as it is, and takes any form of happiness or unhappiness into account. While a utilitarian will no doubt seek to diminish the incidence of feelings which are utilitarianly irrational – or at least of

disagreeable feelings which are so – he might be expected to take them into account while they exist. This is without doubt classical utilitarian doctrine, but there is good reason to think that utilitarianism cannot stick to it without embracing results which are startlingly unacceptable and perhaps self-defeating.

Suppose that there is in a certain society a racial minority. Considering merely the ordinary interests of the other citizens, as opposed to their sentiments, this minority does no particular harm; we may suppose that it does not confer any very great benefits either. Its presence is in those terms neutral or mildly beneficial. However, the other citizens have such prejudices that they find the sight of this group, even the knowledge of its presence, very disagreeable. Proposals are made for removing in some way this minority. If we assume various quite plausible things (as that programmes to change the majority sentiment are likely to be protracted and ineffective) then even if the removal would be unpleasant for the minority, a utilitarian calculation might well end up favouring this step, especially if the minority were a rather small minority and the majority were very severely prejudiced, that is to say, were made very severely uncomfortable by the presence of the minority.

A utilitarian might find that conclusion embarrassing; and not merely because of its nature, but because of the grounds on which it is reached. While a utilitarian might be expected to take into account certain other sorts of consequences of the prejudice, as that a majority prejudice is likely to be displayed in conduct disagreeable to the minority, and so forth, he might be made to wonder whether the unpleasant experiences of the prejudiced people should be allowed, *merely as such*, to count. If he does count them, merely as such, then he has once more separated himself from a body of ordinary moral thought which he might have hoped to accommodate; he may also have started on the path of

defeating his own view of things. For one feature of these sentiments is that they are from the utilitarian point of view itself irrational, and a thoroughly utilitarian person would either not have them, or if he found that he did tend to have them, would himself seek to discount them. Since the sentiments in question are such that a rational utilitarian would discount them in himself, it is reasonable to suppose that he should discount them in his calculations about society; it does seem quite unreasonable for him to give just as much weight to feelings – considered just in themselves, one must recall, as experiences of those that have them – which are essentially based on views which are from a utilitarian point of view irrational, as to those which accord with utilitarian principles. Granted this idea, it seems reasonable for him to rejoin a body of moral thought in other respects congenial to him, and discount those sentiments, just considered in themselves, totally, on the principle that no pains or discomforts are to count in the utilitarian sum which their subjects have just because they hold views which are by utilitarian standards irrational. But if he accepts that, then in the cases we are at present considering no extra weight at all can be put in for bad feelings of George or Jim about their choices, if those choices are, leaving out those feelings, on the first round utilitarianly rational.

The psychological effect on the agent was the first of two general effects considered by utilitarians, which had to be discussed. The second is in general a more substantial item, but it need not take so long, since it is both clearer and has little application to the present cases. This is the *precedent effect*. As Burke rightly emphasized, this effect can be important: that one morally *can* do what someone has actually done, is a psychologically effective principle, if not a deontically valid one. For the effect to operate, obviously some conditions must hold on the publicity of the act and on such things as the status of the agent (such considerations

weighed importantly with Sir Thomas More); what these may be will vary evidently with circumstances.

In order for the precedent effect to make a difference to a utilitarian calculation, it must be based upon a confusion. For suppose that there is an act which would be the best in the circumstances, except that doing it will encourage by precedent other people to do things which will not be the best things to do. Then the situation of those other people must be relevantly different from that of the original agent; if it were not, then in doing the same as what would be the best course for the original agent, they would necessarily do the best thing themselves. But if the situations are in this way relevantly different, it must be a confused perception which takes the first situation, and the agent's course in it, as an adequate precedent for the second.

However, the fact that the precedent effect, if it really makes a difference, is in this sense based on a confusion, does not mean that it is not perfectly real, nor that it is to be discounted: social effects are by their nature confused in this sort of way. What it does emphasize is that calculations of the precedent effect have got to be realistic, involving considerations of how people are actually likely to be influenced. In the present examples, however, it is very implausible to think that the precedent effect could be invoked to make any difference to the calculation. Jim's case is extraordinary enough, and it is hard to imagine who the recipients of the effect might be supposed to be; while George is not in a sufficiently public situation or role for the question to arise in that form, and in any case one might suppose that the motivations of others on such an issue were quite likely to be fixed one way or another already.

No appeal, then, to these other effects is going to make a difference to what the utilitarian will decide about our examples. Let us now look more closely at the structure of those decisions.

5. Integrity

The situations have in common that if the agent does not do a certain disagreeable thing, someone else will, and in Jim's situation at least the result, the state of affairs after the other man has acted, if he does, will be worse than after Jim has acted, if Jim does. The same, on a smaller scale, is true of George's case. I have already suggested that it is inherent in consequentialism that it offers a strong doctrine of negative responsibility: if I know that if I do X, O_1 will eventuate, and if I refrain from doing X, O_2 will, and that O_2 is worse than O_1, then I am responsible for O_2 if I refrain voluntarily from doing X. 'You could have prevented it', as will be said, and truly, to Jim, if he refuses, by the relatives of the other Indians. (I shall leave the important question, which is to the side of the present issue, of the obligations, if any, that nest round the word 'know': how far does one, under utilitarianism, have to research into the possibilities of maximally beneficent action, including prevention?)

In the present cases, the situation of O_2 includes another agent bringing about results worse than O_1. So far as O_2 has been identified up to this point – merely as the worse outcome which will eventuate if I refrain from doing X – we might equally have said that what that other brings about is O_2; but that would be to underdescribe the situation. For what occurs if Jim refrains from action is not solely twenty Indians dead, but *Pedro's killing twenty Indians*, and that is not a result which Pedro brings about, though the death of the Indians is. We can say: what one does is not included in the outcome of what one does, while what another does can be included in the outcome of what one does. For that to be so, as the terms are now being used, only a very weak condition has to be satisfied: for Pedro's killing the Indians to be the outcome of Jim's refusal, it only has to be causally true that if Jim had not refused, Pedro would not have done it.

That may be enough for us to speak, in some sense, of Jim's responsibility for that outcome, if it occurs; but it is certainly not enough, it is worth noticing, for us to speak of Jim's *making* those things happen. For granted this way of their coming about, he could have made them happen only by making Pedro shoot, and there is no acceptable sense in which his refusal makes Pedro shoot. If the captain had said on Jim's refusal, 'you leave me with no alternative', he would have been lying, like most who use that phrase. While the deaths, and the killing, may be the outcome of Jim's refusal, it is misleading to think, in such a case, of Jim having an *effect* on the world through the medium (as it happens) of Pedro's acts; for this is to leave Pedro out of the picture in his essential role of one who has intentions and projects, projects for realizing which Jim's refusal would leave an opportunity. Instead of thinking in terms of supposed effects of Jim's projects on Pedro, it is more revealing to think in terms of the effects of Pedro's projects on Jim's decision. This is the direction from which I want to criticize the notion of negative responsibility.

There are of course other ways in which this notion can be criticized. Many have hoped to discredit it by insisting on the basic moral relevance of the distinction between action and inaction, between intervening and letting things take their course. The distinction is certainly of great moral significance, and indeed it is not easy to think of any moral outlook which could get along without making some use of it. But it is unclear, both in itself and in its moral applications, and the unclarities are of a kind which precisely cause it to give way when, in very difficult cases, weight has to be put on it. There is much to be said in this area, but I doubt whether the sort of dilemma we are considering is going to be resolved by a simple use of this distinction. Again, the issue of negative responsibility can be pressed on the question of how limits are to be placed on one's apparently

boundless obligation, implied by utilitarianism, to improve the world. Some answers are needed to that, too – and answers which stop short of relapsing into the bad faith of supposing that one's responsibilities could be adequately characterized just by appeal to one's roles.[1] But, once again, while that is a real question, it cannot be brought to bear directly on the present kind of case, since it is hard to think of anyone supposing that in Jim's case it would be an adequate response for him to say that it was none of his business.

What projects does a utilitarian agent have? As a utilitarian, he has the general project of bringing about maximally desirable outcomes; how he is to do this at any given moment is a question of what causal levers, so to speak, are at that moment within reach. The desirable outcomes, however, do not just consist of agents carrying out *that* project; there must be other more basic or lower-order projects which he and other agents have, and the desirable outcomes are going to consist, in part, of the maximally harmonious realization of those projects ('in part', because one component of a utilitarianly desirable outcome may be the occurrence of agreeable experiences which are not the satisfaction of anybody's projects). Unless there were first-order projects, the general utilitarian project would have nothing to work on, and would be vacuous. What do the more basic or lower-order projects comprise? Many will be the obvious kinds of desires for things for oneself, one's family, one's friends, including basic necessities of life, and in more relaxed circumstances, objects of taste. Or there may be pursuits and interests of an intellectual, cultural or creative character. I introduce those as a separate class not because the objects of them lie in a separate class, and provide – as some utilitarians, in their churchy way, are fond of saying – 'higher' pleasures. I introduce them separately because the agent's identification

[1] For some remarks bearing on this, see *Morality*, the section on 'Goodness and roles', and Cohen's article there cited.

with them may be of a different order. It does not have to be: cultural and aesthetic interests just belong, for many, along with any other taste; but some people's commitment to these kinds of interests just is at once more thoroughgoing and serious than their pursuit of various objects of taste, while it is more individual and permeated with character than the desire for the necessities of life.

Beyond these, someone may have projects connected with his support of some cause: Zionism, for instance, or the abolition of chemical and biological warfare. Or there may be projects which flow from some more general disposition towards human conduct and character, such as a hatred of injustice, or of cruelty, or of killing.

It may be said that this last sort of disposition and its associated project do not count as (logically) 'lower-order' relative to the higher-order project of maximizing desirable outcomes; rather, it may be said, it is itself a 'higher-order' project. The vital question is not, however, how it is to be classified, but whether it and similar projects are to count among the projects whose satisfaction is to be included in the maximizing sum, and, correspondingly, as contributing to the agent's happiness. If the utilitarian says 'no' to that, then he is almost certainly committed to a version of utilitarianism as absurdly superficial and shallow as Benthamite versions have often been accused of being. For this project will be discounted, presumably, on the ground that it involves, in the specification of its object, the mention of other people's happiness or interests: thus it is the kind of project which (unlike the pursuit of food for myself) presupposes a reference to other people's projects. But that criterion would eliminate any desire at all which was not blankly and in the most straightforward sense egoistic.[1] Thus we should be reduced

[1] On the subject of egoistic and non-egoistic desires, see 'Egoism and altruism', in *Problems of the Self* (Cambridge University Press, London, 1973).

to frankly egoistic first-order projects, and – for all essential purposes – the one second-order utilitarian project of maximally satisfying first-order projects. Utilitarianism has a tendency to slide in this direction, and to leave a vast hole in the range of human desires, between egoistic inclinations and necessities at one end, and impersonally benevolent happiness-management at the other. But the utilitarianism which has to leave this hole is the most primitive form, which offers a quite rudimentary account of desire. Modern versions of the theory are supposed to be neutral with regard to what sorts of things make people happy or what their projects are. Utilitarianism would do well then to acknowledge the evident fact that among the things that make people happy is not only making other people happy, but being taken up or involved in any of a vast range of projects, or – if we waive the evangelical and moralizing associations of the word – commitments. One can be committed to such things as a person, a cause, an institution, a career, one's own genius, or the pursuit of danger.

Now none of these is itself the *pursuit of happiness*: by an exceedingly ancient platitude, it is not at all clear that there could be anything which was just that, or at least anything that had the slightest chance of being successful. Happiness, rather, requires being involved in, or at least content with, something else.[1] It is not impossible for utilitarianism to accept that point: it does not have to be saddled with a naïve and absurd philosophy of mind about the relation between desire and happiness. What it does have to say is that if such

[1] This does not imply that there is no such thing as the project of pursuing pleasure. Some writers who have correctly resisted the view that all desires are desires for pleasure, have given an account of pleasure so thoroughly adverbial as to leave it quite unclear how there could be a distinctively hedonist way of life at all. Some room has to be left for that, though there are important difficulties both in defining it and living it. Thus (particularly in the case of the very rich) it often has highly ritual aspects, apparently part of a strategy to counter boredom.

commitments are worth while, then pursuing the projects that flow from them, and realizing some of those projects, will make the person for whom they are worth while, happy. It may be that to claim that is still wrong: it may well be that a commitment can make sense to a man (can make sense of his life) without his supposing that it will make him *happy*.[1] But that is not the present point; let us grant to utilitarianism that all worthwhile human projects must conduce, one way or another, to happiness. The point is that even if that is true, it does not follow, nor could it possibly be true, that those projects are themselves projects of pursuing happiness. One has to believe in, or at least want, or quite minimally, be content with, other things, for there to be anywhere that happiness can come from.

Utilitarianism, then, should be willing to agree that its general aim of maximizing happiness does not imply that what everyone is doing is just pursuing happiness. On the contrary, people have to be pursuing other things. What those other things may be, utilitarianism, sticking to its professed empirical stance, should be prepared just to find out. No doubt some possible projects it will want to discourage, on the grounds that their being pursued involves a negative balance of happiness to others: though even there, the unblinking accountant's eye of the strict utilitarian will have something to put in the positive column, the satisfactions of the destructive agent. Beyond that, there will be a vast variety of generally beneficent or at least harmless projects; and some no doubt, will take the form not just of tastes or fancies, but of what I have called 'commitments'. It may even be that the utilitarian researcher will find that many of those with commitments, who have really identified themselves with objects outside themselves, who are thoroughly involved with other persons, or institutions, or activities or

[1] For some remarks on this possibility, see *Morality*, section on 'What is morality about?'

causes, are actually happier than those whose projects and wants are not like that. If so, that is an important piece of utilitarian empirical lore.

When I say 'happier' here, I have in mind the sort of consideration which any utilitarian would be committed to accepting: as for instance that such people are less likely to have a break-down or commit suicide. Of course that is not all that is actually involved, but the point in this argument is to use to the maximum degree utilitarian notions, in order to locate a breaking point in utilitarian thought. In appealing to this strictly utilitarian notion, I am being more consistent with utilitarianism than Smart is. In his struggles with the problem of the brain-electrode man, Smart (p. 22) commends the idea that 'happy' is a partly evaluative term, in the sense that we call 'happiness' those kinds of satisfaction which, as things are, we approve of. But *by what standard* is this surplus element of approval supposed, from a utilitarian point of view, to be allocated? There is no source for it, on a strictly utilitarian view, except further degrees of satisfaction, but there are none of those available, or the problem would not arise. Nor does it help to appeal to the fact that we dislike in prospect things which we like when we get there, for from a utilitarian point of view it would seem that the original dislike was merely irrational or based on an error. Smart's argument at this point seems to be embarrassed by a well-known utilitarian uneasiness, which comes from a feeling that it is not respectable to ignore the 'deep', while not having anywhere left in human life to locate it.[1]

Let us now go back to the agent as utilitarian, and his higher-order project of maximizing desirable outcomes. At this level, he is committed only to that: what the outcome will actually consist of will depend entirely on the facts, on

[1] One of many resemblances in spirit between utilitarianism and high-minded evangelical Christianity.

what persons with what projects and what potential satisfactions there are within calculable reach of the causal levers near which he finds himself. His own substantial projects and commitments come into it, but only as one lot among others – they potentially provide one set of satisfactions among those which he may be able to assist from where he happens to be. He is the agent of the satisfaction system who happens to be at a particular point at a particular time: in Jim's case, our man in South America. His own decisions as a utilitarian agent are a function of all the satisfactions which he can affect from where he is: and this means that the projects of others, to an indeterminately great extent, determine his decision.

This may be so either positively or negatively. It will be so positively if agents within the causal field of his decision have projects which are at any rate harmless, and so should be assisted. It will equally be so, but negatively, if there is an agent within the causal field whose projects are harmful, and have to be frustrated to maximize desirable outcomes. So it is with Jim and the soldier Pedro. On the utilitarian view, the undesirable projects of other people as much determine, in this negative way, one's decisions as the desirable ones do positively: if those people were not there, or had different projects, the causal nexus would be different, and it is the actual state of the causal nexus which determines the decision. The determination to an indefinite degree of my decisions by other people's projects is just another aspect of my unlimited responsibility to act for the best in a causal framework formed to a considerable extent by their projects.

The decision so determined is, for utilitarianism, the right decision. But what if it conflicts with some project of mine? This, the utilitarian will say, has already been dealt with: the satisfaction to you of fulfilling your project, and any satisfactions to others of your so doing, have already been

through the calculating device and have been found inadequate. Now in the case of many sorts of projects, that is a perfectly reasonable sort of answer. But in the case of projects of the sort I have called 'commitments', those with which one is more deeply and extensively involved and identified, this cannot just by itself be an adequate answer, and there may be no adequate answer at all. For, to take the extreme sort of case, how can a man, as a utilitarian agent, come to regard as one satisfaction among others, and a dispensable one, a project or attitude round which he has built his life, just because someone else's projects have so structured the causal scene that that is how the utilitarian sum comes out?

The point here is not, as utilitarians may hasten to say, that if the project or attitude is that central to his life, then to abandon it will be very disagreeable to him and great loss of utility will be involved. I have already argued in section 4 that it is not like that; on the contrary, once he is prepared to look at it like that, the argument in any serious case is over anyway. The point is that he is identified with his actions as flowing from projects and attitudes which in some cases he takes seriously at the deepest level, as what his life is about (or, in some cases, this section of his life – seriousness is not necessarily the same as persistence). It is absurd to demand of such a man, when the sums come in from the utility network which the projects of others have in part determined, that he should just step aside from his own project and decision and acknowledge the decision which utilitarian calculation requires. It is to alienate him in a real sense from his actions and the source of his action in his own convictions. It is to make him into a channel between the input of everyone's projects, including his own, and an output of optimific decision; but this is to neglect the extent to which *his* actions and *his* decisions have to be seen as the actions and decisions which flow from the projects and attitudes with

which he is most closely identified. It is thus, in the most literal sense, an attack on his integrity.[1]

These sorts of considerations do not in themselves give solutions to practical dilemmas such as those provided by our examples; but I hope they help to provide other ways of thinking about them. In fact, it is not hard to see that in George's case, viewed from this perspective, the utilitarian solution would be wrong. Jim's case is different, and harder. But if (as I suppose) the utilitarian is probably right in this case, that is not to be found out just by asking the utilitarian's questions. Discussions of it – and I am not going to try to carry it further here – will have to take seriously the distinction between my killing someone, and its coming about because of what I do that someone else kills them: a distinction based, not so much on the distinction between action and inaction, as on the distinction between my projects and someone else's projects. At least it will have to start by taking that seriously, as utilitarianism does not; but then it will have to build out from there by asking why that distinction seems to have less, or a different, force in this case than it has in George's. One question here would be how far one's powerful objection to killing people just is, in fact, an application of a powerful objection to their being killed. Another dimension of that is the issue of how much it matters that the people at risk are actual, and there, as opposed to hypothetical, or future, or merely elsewhere.[2]

[1] Interestingly related to these notions is the Socratic idea that courage is a virtue particularly connected with keeping a clear sense of what one regards as most important. They also centrally raise questions about the value of pride. Humility, as something beyond the real demand of correct self-appraisal, was specially a Christian virtue because it involved subservience to God. In a secular context it can only represent subservience to other men and their projects.

[2] For a more general discussion of this issue see Charles Fried, *An Anatomy of Values* (Harvard University Press, Cambridge, Mass., 1970), Part Three.

There are many other considerations that could come into such a question, but the immediate point of all this is to draw one particular contrast with utilitarianism: that to reach a grounded decision in such a case should not be regarded as a matter of just discounting one's reactions, impulses and deeply held projects in the face of the pattern of utilities, nor yet merely adding them in – but in the first instance of trying to understand them.

Of course, time and circumstances are unlikely to make a grounded decision, in Jim's case at least, possible. It might not even be decent. Instead of thinking in a rational and systematic way either about utilities or about the value of human life, the relevance of the people at risk being present, and so forth, the presence of the people at risk may just have its effect. The significance of the immediate should not be underestimated. Philosophers, not only utilitarian ones, repeatedly urge one to view the world *sub specie aeternitatis*,[1] but for most human purposes that is not a good *species* to view it under. If we are not agents of the universal satisfaction system, we are not primarily janitors of any system of values, even our own: very often, we just act, as a possibly confused result of the situation in which we are engaged. That, I suspect, is very often an exceedingly good thing. To what extent utilitarians regard it as a good thing is an obscure question. To that sort of question I now turn.

6. The indirect pursuit of utility

Smart's defence is devoted to act-utilitarianism, which (taking for granted the complications which we have pursued in section 2) stands as the view that the rightness of any particular act depends on the goodness of its conse-

[1] Cf. Smart, p. 63.

quences. This is what I called in section 1 *direct* consequentialism; where the goodness of the consequences is cashed in terms of happiness, we can speak of direct utilitarianism. What is direct utilitarianism contrasted with? We cannot just say that direct utilitarianism considers only the utility of actions, while indirect utilitarianism, by contrast, is prepared to consider the utility of things other than actions, such as rules, institutions and dispositions of character. Clearly the act-utilitarian must be prepared to consider the utility of anything: his aim is to maximize utility, and anything, of whatever type, whose existence, introduction or whatever has effects on the amount of satisfaction in the world must be a candidate for assessment by the utilitarian standard. Thus if there *is* anything which has got a utility which cannot be counted in terms of the utility of particular acts, then the utility of that thing as well must be of interest to the direct utilitarian.

Here someone might say that there was nothing which had a utility which could not be counted in terms of the utility of particular acts. If institutions or rules or dispositions of character possess utility, then they possess it in terms of the acts which they variously encourage, license, enjoin or lead to. To take, in particular, the case of a rule: consider two states of society, one in which a given rule obtains, and another in which it does not. If there is a difference of utility between them which relates to this difference, then (it may be said) there must be a difference in the set of acts which occur in the two states, to which that difference in utility must be traceable. Different acts are done as a result of the rule obtaining. There have to be some such acts, on anyone's view, if we are to say that the rule *obtains* at all; other acts come into it in terms of rules being inculcated, thought of, brought up as matters of reproach, and in many other ways. In the end, it may be said, the total utility effect of a rule's obtaining must be cashable in terms of the effects of acts.

Let us call this, in a barbarous phrase, the 'act-adequacy premiss'.[1]

But if that premiss is right, then it becomes unclear what the difference between direct and indirect utilitarianism is. For so long at least as we regard utilitarianism as a system of *total assessment* – as providing an answer, basically, to the question 'how is the world going?' – then, on the present argument, it looks as though anything that anybody else can do, the direct utilitarian can do at least as well. If all the other candidates for utilitarian assessment, such as rules, can have their differential utility cashed in terms of acts, then the direct utilitarian can assess their contribution to the world as well as he can assess acts which are not particularly associated with rules. Whether the total utility of the social state in which the rule obtains is greater or less, measured by these means, than that of a state in which it does not obtain, then appears to be a totally empirical question, and it can scarcely be that the difference between direct and indirect utilitarianism consists just in giving different answers to that.

Suppose, on the other hand, that the act-adequacy premiss is false, and that there is, as it were, a surplus causal effect of the rule's obtaining which cannot be expressed in terms of the effects of acts. Then indeed, the direct utilitarian will not be able to capture all differences of utility just by counting the utility of acts. But equally, if that is so, then he must, from a utilitarian point of view, be quite irrational in insisting on so doing. As a utilitarian, as we said just now, he must be concerned with the utility in the world – and if utility can leak into or out of the world by channels which do not run totally through acts, then he would be mad to take no account of them.

As systems of total assessment then, it looks so far as though

[1] I shall not try to fill in any more determinate content for this premiss; its role in the following arguments is of an essentially formal character.

either there really is no difference between direct and indirect utilitarianism, in the sense that the direct utilitarian can also take into account the effects of such things as rules, and it is just an empirical question what the effects are; or else there is a difference, and direct utilitarianism must, in terms of the overall aims of utilitarianism, be irrational, simply because it would be ignoring important sources of utility. Thus a large question seems to have been rather rapidly short-circuited. The reason for this is that we have started too far out, as it were, by comparing the two outlooks *as systems of total assessment*, and by asking whether, as such systems, they were concerned with anything but acts. Rather, we have to start by asking, to the extent that they are both concerned with acts, *how* they are each concerned with them.

The place to start, then, is back with the assessment of acts themselves. Thus, as we said at the beginning, direct utilitarianism regards that act as right which has the best consequences. So indirect utilitarianism may be expected to deny this, and to hold that some acts are right even though they are not utility-maximizing – for instance, because they are done in accordance with a rule which is utilitarianly valuable. Another version might be that an act could be right just because it was the expression of a character-disposition, the obtaining of which in society is utilitarianly valuable. Thus the difference might be captured in some such way, at the level of the assessment of particular acts. We must remember, however, that it is precisely with regard to the rightness of the acts, and not necessarily elsewhere in the contrast just sketched, that the difference can be captured. Thus if we ask the indirect utilitarian "What does the rightness of acts consist in?" we shall get an answer with which the direct utilitarian will to a certain extent disagree; and conversely. But if we ask either of them "In what does the value of rules, traits of character, etc., consist?", we shall not necessarily get different answers. For

if the indirect utilitarian accepts what I called earlier the 'act-adequacy premiss' – and there is no inconsistency in his doing so – then he will reply "in the utility of the acts that follow on their existence", which is the answer that the direct utilitarian will give as well.

Not only can they agree on that, under the act-adequacy assumption, but they can agree importantly on its interpretation. Thus, to confine ourselves to the case of rules, they can agree, if they are sensible, that the utility of acts that follow on the obtaining of a rule is not to be equated with the utility of acts that consist in obeying the rule.[1] If the rule actually *obtains* in society, as opposed to having merely been promulgated, e.g. from some legal source, then we can say – by definition, indeed – that a good number of acts that are performed because it exists must be acts of obedience to it. But very many other acts, of many different kinds, are performed because a rule exists. Some of them we touched on earlier; they range from acts involved in teaching it to acts involved in avoiding detection for breaking it, and all make their contribution to the overall utility of its obtaining. To equate the utility of a rule's obtaining with the utility of its being followed is not the mark of any utilitarian doctrine, direct or indirect – it is just a sign of simple-mindedness.

Thus the distinction has turned out to centre on the rightness or wrongness of particular acts. But now the question arises of how the distinction, so set up, is to be used, and of what follows from particular acts being right or wrong for these different reasons. In particular, I shall ask these questions of Smart; other utilitarians, perhaps of more cognitivist outlook, may give different answers, but to

[1] Cf. Lyons's distinction between 'acceptance-utility' and 'following-utility': *The Forms and Limits of Utilitarianism* (Oxford University Press, London, 1965), pp. 137 ff. Further distinctions are needed when there is question of formally adopting or promulgating a rule – thus it may not be obeyed at all. But these need not concern us here.

some extent these problems will arise with any current version of utilitarianism. On Smart's view, one thing certainly is *not* meant by saying that an act is right if it maximizes utility – namely, that if the act maximizes utility, then it will be right to announce to the agent himself, or to anyone else, that the act is right. For any such announcements must fall under the provision that he makes about praise and blame,[1] where the only consideration is the effectiveness or utility of the utterance, and that does not, as he several times reminds us, necessarily come to the same thing as the utility of the act which the utterance relates to. Thus he encourages the patient utilitarian faced with the magical society to think it better to commend and blame acts by the local standard rather than a utilitarian one, since confusion and disutility are likely to follow from an ill-considered dash by the natives to accommodate themselves to the utterances of this influential commentator.[2]

Smart's causal theory of moral comment has two familiar disadvantages. One is that, as a practice, it essentially lacks openness – that is to say, it is not possible for it to be openly known in the society what this practice is. If it were known, then it would in some part cease to work, since one important dimension, at least, by which moral comment can be efficacious is by those who receive it not thinking in terms of its being efficacious or not, but in terms of whether it is justified. It is a very evident fact that blame has a decreasing, or a counter-productive, effect if it is handed out in ways which its objects perceive as unfair. In regarding it as fair or not, its objects cannot merely be considering whether it will work or not. Thus if those who administer the blame, or some smaller class of knowing utilitarians standing behind those who administer it, do in effect think of the question of fairness as fundamentally the same as the question of efficacy,

[1] P. 53 al.

[2] Smart, p. 50.

then there has to be disingenuousness between them and the others, and the institution has to lack openness, in the sense that it will not work as an institution unless there is widespread ignorance about its real nature. This lack of openness, a notable feature of the arrangements which Smart proposes, I shall come back to in section 7.

The second weakness of the causal theory of moral comment is that it makes it very difficult to make sense of a man's view of his own conduct; particularly if he himself believes the causal theory, since then the lack of openness I have mentioned stands between the man and himself – it is hard to see how he can blame himself if he knows what he is doing in doing that. Now utilitarians in fact are not very keen on people blaming themselves, which they see as an unproductive activity: not to cry over spilt milk figures prominently among utilitarian proverbial injunctions (and carries the characteristically utilitarian thought that anything you might want to cry over is, like milk, replaceable). Rather, they are concerned with practical future effects, and the question of what is the right thing to do focusses essentially on the situation of decision: the central question is not, "did he (or I) do the right thing?" but "what is the right thing to do?". This emphasis on the practical decision-making aspect of moral thought is of course not peculiar to utilitarianism, but it is not surprising that utilitarianism should particularly emphasize it.

If the central question is the practical question of what is the right thing to do, the problem now is, what distinctive contribution to understanding that question and answering it does the direct utilitarian give us? He tells us that the answer to the question "what is the right thing to do?" is to be found in that act which has the best consequences. But it seems difficult to put that to any use in this connexion, except by taking it to imply the following: that the correct question to ask, if asking what is the right thing to do, is

what act will have the best consequences. But the moment that has been accepted, we lose a distinction on which Smart, following Sidgwick, lays great weight – that between justification and motivation.

Smart makes much of this distinction, to reject the immediately calculative aspect of utilitarianism, and to commend such things as spontaneity[1], and he even is prepared to consider, though he rejects, Moore's idea that an act-utilitarian might never act in the spirit of an act- utilitarian.[2] But now, if our argument is right, it looks as though Smart has no room, or at least very little room, in which to make these manoeuvres. For we have tracked down the distinction between direct utilitarianism and other sorts of utilitarianism to a difference about what acts are right; and we have located the significance of that question, for a utilitarian, in the situation of decision; and we have found no alternative to taking its significance in that situation as a matter of the correct question to ask oneself; and that makes it a matter of motivation, of what people should think about in deciding what to do. So if Smart wishes to sustain a distinctively direct utilitarian position, then he cannot also use some of the devices of indirect utilitarianism to take the edge off it.

There is one area in which Smart himself seems happy to accept that point, namely with regard to rules. He says that if a utilitarian agent perceives that in particular circumstances the course with the best consequences all round consists in breaking the rule, then it would be 'rule-worship' not to do so; and that a utilitarian should regard rules as 'rules of thumb'[3]. I interpret this in the light of his remark that the primary idea of having rules is to save time[4]. There is indeed

[1] Pp. 44–5.
[2] Pp. 43–4.
[3] P. 42.
[4] Or to deal with cases where there is no time: p. 42.

a coherent model of that sort of rule, which I have elsewhere[1] called the 'gas bill' model, which refers to the situation in which the cost to an enterprise of interfering with a fixed process for handling transactions and halting a given item, is greater than a loss which is indeed incurred on that item. That model makes it clear why, for an individual, the value of rules of thumb is costed principally in terms of time. It also illuminates the point that once an agent has perceived the disutility in the particular case, there is no point in his following the rule in that case; for coming to perceive the immediate disutility is the individual analogy to interrupting the commercial process – the reflective intervention which costs the time has already been made.

There are of course cases in which following a 'rule of thumb' will generate more disutility than breaking it. But necessarily, of course, there is no certain way of identifying such cases in advance: for to make sure of each case whether it was or was not of that sort would involve in each case the reflective intervention which it is the point of the rule to avoid. So anyone who adopts a 'rule of thumb' will know in advance that there will be some exceptional cases which will not announce themselves as exceptional cases; that is, he will know that he is licensing some tactical disutility in the pursuit of strategic utility. Now, if the facts are as we have supposed, he will not be able to avoid losing some utility, since the alternative is to consider every case, and considering every case has, in sum, greater disutility. But he will know in advance that some of the actions he will do will not be, by direct utilitarian standards, the right actions, or even, relative to the evidence one could have gathered if one had investigated the particular cases, probably the right actions. Thus there will be a utilitarian type of reason for thinking it better to adopt a course of action which

[1] *Morality*, section on utilitarianism.

involves, one already knows, not always doing the right action.

To accept this last point does not involve abandoning what I earlier called the 'act-adequacy premiss'. One could accept the last point, and still think that all utility-changes in the world were induced via actions; one would merely have to recognize that one's sometimes doing wrong actions was a necessary condition of more optimific actions being done, even by oneself. This is the sort of spirit, perhaps, in which Smart suggests that knocking off good works for a bit might be a means to doing more good works.[1] In fact, I think that that is as far as Smart is prepared to take it, at least when he is thinking strictly in terms of direct utilitarianism as a personal morality: rules of thumb will be acceptable to me in so far as they render it more probable that they will lead to more right actions in the long run being done by me. Moreover, if they are to have that tendency, it is important that I *treat* them as rules of thumb, which means not only that if I do discover that this is an exceptional case, then I treat it as an exception, but also – and importantly – I keep a utilitarian eye open for signs that a case may be exceptional.

But if these precautions are rational, then clearly the utilitarian agent had better not go too far in the direction of cultivating spontaneity or a lack of conscious concern for utilitarian considerations, since every step in that direction must tend to decrease the probability that he will do right actions; unless one believes either that the Invisible Hand of early capitalism will guide the unreflective agent to utilitarianly desirable outcomes, or else that rationally utilitarian deliberation in particular cases is actually harmful to utilitarian outcomes *in those cases* (even apart from loss of time, etc.), which may well be true, but can hardly be believed

[1] P. 55.

by, at least, a direct utilitarian. It is for these reasons, no doubt, that while Smart does make some excursions into licensing non-utilitarian states of mind, he displays some caution in doing so. The relaxing from good works not only will, it is hoped, produce more good works, but is designed to; and if spontaneity has utilitarian value, then doubtless we can organize some spontaneity. That Smart's direct utilitarianism is in fact cautious about commending dispositions which are psychologically removed from the calculation of utilities is suggested also by his saying virtually nothing about excellencies of character which might go into the specification of a good man, or various sorts of good man; and that his account of that notion itself is done entirely in terms of a man's maximizing right actions.[1]

It is consistent of Smart, I believe, to restrict departures from utilitarian calculation, if he is going to be a direct utilitarian; but then it is not consistent of him to present direct utilitarianism as a doctrine merely about justification and not about motivation. There is no distinctive place for *direct* utilitarianism unless it is, within fairly narrow limits, a doctrine about how one should decide what to do. This is because its distinctive doctrine is about what acts are right, and, especially for utilitarians, the only distinctive interest or point of the question what acts are right, relates to the situation of deciding to do them.

In one, and the most obvious, way, direct utilitarianism is the paradigm of utilitarianism – it seems, in its blunt insistence on maximizing utility and its refusal to fall back on rules and so forth, of all utilitarian doctrines the most faithful

[1] 'A good agent is one who acts more nearly in a generally optimific way that does the average one' (p. 48). It is not in the least clear what this means, but it does seem to represent a rather relaxed standard: thus the well-known difficulty of finding ten good men in Sodom (Genesis 18–19) should perhaps not have arisen, unless Sodom had an exceedingly small population.

to the spirit of utilitarianism, and to its demand for a rational, decidable, empirically based, and unmysterious set of values. At the same time, however, it contains something which a utilitarian could see as a certain weakness, a traditional idea which it unreflectively harbours. This is, that the best world must be one in which right action is maximized. Under utilitarianism, it is not clear that this claim even has to be true; and when it is true, it turns out more trivial than it looks.

If the act-adequacy premiss is false, the claim need not even be true. Imagine that the greatest utility was in fact produced by people displaying and witnessing spontaneous and zestful activity. Many particular acts would be wrong, in the sense that if these acts were replaced there could be an increase in utility; but there is no way of replacing them without destroying the spontaneity and zest. Here right acts are sacrificed, indeed to greater utility, but not to greater utility which involves any larger number of right *acts* – it lies rather in a certain style and spirit of action. If, on the other hand, the act-adequacy premiss is true, then right action should be maximized, since what will be bought by a system which involves individually wrong acts will be, in this case, a larger number of right acts. But this is a triviality. For even if right acts were being maximized; and even if, further, my act were individually necessary to that being so, so that even this act of mine were, selectively, right: it would not follow that its utilitarian rightness would be evident to me or to anyone else in the situation.

An example, boringly fanciful and schematic in itself, may illustrate the point here. A utilitarian enlightened community might find that there was a tendency among the citizens to slip away from the utilitarian spirit, making reckless decisions themselves, and grumbling about arrangements which scientific enquiry had shown to be for the best. The most painless way of curing this is to find a means to remind them

of the disadvantages of not being utilitarian. The government establishes a reservation of profoundly non-utilitarian persons, of Old Testament or other magical persuasion, leaves them to get on with their lives, and by secret means transmits by TV to the rest of the people some of the more richly counter-utilitarian consequences of their way of life. If this worked as planned, the non-utilitarian acts of those in the reservation would in fact be utilitarianly right, or at least some indeterminably large proportion of them would be (the allocation of marginal effects would be impossible); but the way in which they in fact contributed to maximizing utility would be one which required almost everyone outside the reservation to regard them as wrong, and those inside the reservation to regard them as right for reasons which for the utilitarian would make them wrong. Thus even granted the act-adequacy premiss, there is nothing but a triviality in the proposition that right acts should be maximized. It does not follow that one should maximize what seem to utilitarians right acts. It may well be best to secure many of what utilitarians will be bound to regard as wrong acts, and there is no reason why the distribution of these between persons should be equal: as the model illustrates, there might be utilitarian reasons for there being a corner in 'wrong' acts among some particular men. Utilitarianism has no more reason to insist on equity in this respect than in any other.

Once one has moved back in this way to the 'total assessment' position, the utility of anything is open to question, including, of course, that of utilitarian thinking as a personal and social phenomenon. There are some powerful reasons for thinking that its prevalence could be a disaster. Some of these are hinted at occasionally by Smart, at those points at which he wishes (as I have suggested, inconsistently) to keep direct utilitarianism and at the same time spirit away utilitarian calculation. Let me mention two others.

First, many of the qualities that human beings prize in society and in one another are notably non-utilitarian, both in the cast of mind that they involve and in the actions they are disposed to produce. There is every reason to suppose that people's *happiness* is linked in various ways to these qualities. It is no good the utilitarian saying that such happiness does not count. For as we have already seen in this connexion, modern utilitarianism is supposed to be a system neutral between the preferences that people actually have, and here are some preferences which some people actually have. To legislate them out is not to pursue people's happiness, but to remodel the world towards forms of 'happiness' more amenable to utilitarian ways of thought. But if they are not to be legislated out, then utilitarianism has got to coexist with them, and it is not clear how it does that. As we have already seen with Smart's remarks on spontaneity, you cannot both genuinely possess this kind of quality and also reassure yourself that while it is free and creative and uncalculative, it is also acting for the best. Here we have that same problem of alienation from one's projects which we considered before in relation to integrity.

Second, there is the *Gresham's Law* problem, related to the well-known problem of games theory, the Prisoner's Dilemma.[1] The upshot of the Dilemma (the details of which need not concern us here) is that it can be individually rational for two players in a competitive game to adopt strategies which jointly produce an outcome worse than could have been achieved by their each adopting another strategy; but while they can both see this, neither of them can afford to adopt the different strategy, for fear that he will do so alone, something which would produce a worse outcome for him (though better for his opponent) than any

[1] For a discussion of the Dilemma, see e.g. Luce and Raiffa, *Games and Decisions* (Wiley, London, 1967). The present argument is a slightly expanded version of one in *Morality*, *loc. cit.*

other. The way out of this is co-operating; one way to that is an 'enforceable agreement', where this can be Hobbesianly interpreted as an agreement with an indefinitely large penalty attached to breaking it. The Dilemma is usually interpreted in terms of self-interested preferences, but a similar structure can arise in a competition between utilitarian agents on one side, and self-interested (or merely opposed) agents on the other. Now society cannot exist without some degree of co-operative and (in Smart's term) benevolent motivation, to some degree internalized, to some degree sustained by sanctions. But the system cannot and does not guarantee peace, both because there are agents who are uncooperative, and also because there are conflicts of view about what may constitute happiness (the utilitarian assumption that it must be possible, by a maximizing function, to combine in some sort of compromise as many people as possible getting as much as possible, just depends on the usual assumptions about the demure and essentially domestic character of what people want).

Once such conflicts cannot be resolved within the usual framework of compromise, utilitarianism has a particular tendency to raise the conflict to new levels. For it must always be the utilitarian's business, thinking as a utilitarian, to take the least bad action necessary and sufficient to prevent the worst outcome: pre-emptive action is of the essence of utilitarian rationality. But since an opponent may know that the utilitarian is a utilitarian and is committed to this, he himself will raise his bid. Both may see, as in the Dilemma, that the joint outcome of these procedures will be very bad, but there is no way in which the utilitarian can cut off the process without taking an unjustified risk with the utilities he is supposed to be maximizing. Thus he is driven on by utilitarian rationality itself to outbid the opponent, and the cumulative process is disastrous, although no particular departure from it can be justified.

Of course the situations in which such conflict can grow are in various ways restricted, in particular within the state, since it is the aim, and if Hobbes is right the function, of state power to contain such conflicts. But there are inter-state conflicts, and conflicts between state power and other forces, and indeed the same structure can apply to conflicts within the state, even if state power suffices to stop the full menu of violent means being explored by the combatants. Moreover, the mere existence of state power is inadequate to contain conflict unless people in the state are to some degree motivated to avoid conflict. Both in the provision of such a motivation, and in the business of limiting potentially limitless conflict, there is reason to think that a distinctively non-utilitarian disposition is needed: a disposition to limit one's reactions, even though in the particular case the cost of so doing may turn out to be high. That is to say, people have to be motivated, and deeply motivated, not to take the means necessary and sufficient to prevent, in the particular case, the worst outcome. A system of dispositions against pre-emptive action – even in the face of strong provocation to utilitarian conduct – has a chance of limiting conflict, and such a system requires people to be brought up and fortified in dispositions not to think of situations in a utilitarian way. This is not to say that they do not think at all in terms of the consequences of their actions – that would be merely insane, if intelligible at all. Nor does it mean they fix one definite limit to their response whatever may threaten, as pacifists do: that would be to suppose that the only alternative to utilitarianism was accepting that there were certain things obligatory whatever the consequences, a position we rejected a long time ago, in section 2. It means rather that the response falls short of what would be utilitarianly required at a given point: and falls firmly and reliably short of it.

Two utilitarian answers can be considered here. The

utilitarian may say, first, that anyone can talk about what would be desirable to limit conflict; no doubt if these dispositions were general, conflict would be contained. But equally, if utilitarianism were general, conflict would be contained. This reply just misses the point of the argument. Let us concede that if utilitarianism were general, conflict would be contained; though in fact, there is some doubt about this, unless one adds that not only do the parties agree on the formalities of utilitarianism, but they share a common, or at least only trivially various, concept of happiness. The point concerns the situation in which not all the parties have co-operative dispositions – that is to say, the actual situation. If one party to a conflict lacks co-operative dispositions, and the other is a strict utilitarian, then the ground is rich for conflict to grow pre-emptively; if the more socialized party has a disposition to resist pre-emption, it may not.

Another utilitarian answer will be that the arguments I have advanced for these dispositions are anyway utilitarian arguments. In a way, that is right, and they are meant to be – they are meant to use utilitarian terms to the maximum degree. But what they show, if correct, is that granted some empirical generalities of a kind which are the background to all problems of morality, utilitarianism's fate is to usher itself from the scene. As we have seen, direct utilitarianism represents certainly a distinctive way of deciding moral questions, a way, however, which there is good reason to think, if generally employed, could lead to disaster; and some qualifications which Smart is disposed to put in seem to signal some recognition of that, and a comprehensible desire to leave the way open for utilitarianism to retire to a more indirect level, towards the dimension of total assessment. But once that has started, there seems nothing to stop, and a lot to encourage, a movement by which it retires to the totally transcendental standpoint from which all it demands is that the world should be ordered for the best, and

that those dispositions and habits of thought should exist in the world which are for the best, leaving it entirely open whether those are themselves of a distinctively utilitarian kind or not. If utilitarianism indeed gets to this point, and determines nothing of how thought in the world is conducted, demanding merely that the way in which it is conducted must be for the best, then I hold that utilitarianism has disappeared, and that the residual position is not worth calling utilitarianism.[1]

If utility could be globally put together at all – and that has been an assumption of these arguments, though I shall raise some doubts about it in the next section – then there might be maximal total utility from the transcendental standpoint, even though nobody in the world accepted utilitarianism at all. Moreover, if the previous arguments have been correct, it is reasonable to suppose that maximal total utility actually requires that few, if any, accept utilitarianism. If that is right, and utilitarianism has to vanish from making any distinctive mark in the world, being left only with the total assessment from the transcendental standpoint – then I leave it for discussion whether that shows that utilitarianism is unacceptable, or merely that no one ought to accept it.

7. Social choice

The fathers of utilitarianism thought of it principally as a system of social and political decision, as offering a criterion and basis of judgement for legislators and administrators. This is recognizably a different matter from utilitarianism as a system of personal morality, but it is hard for a number of important reasons to keep the two things ultimately apart, and to stop the spirit of utilitarianism, firmly established in

[1] For a similar view, cf. John Rawls, *A Theory of Justice* (Oxford University Press, London, 1972), pp. 182, 184–5.

one, from moving into the other. If individual decisions on personal matters are made on a utilitarian basis, then those citizens will both direct the same outlook on to their views about what should be done in the public sphere, and also expect the legislature and the executive to make its decisions in that spirit. Indeed, a utilitarian is likely to think that the case for public utilitarianism is even stronger than that for private. For one thing, the decisions of government[1] affect more persons, in the main, than private decisions. But, more than that, he is likely to feel that there is something in the nature of modern government (at least) which requires the utilitarian spirit. Private citizens might legitimately, if regrettably, have religious beliefs or counter-utilitarian ideals, but government in a secular state must be secular, and must use a system of decision which is minimally committed beyond its intrinsic commitment to the welfare of its citizens. Thus utilitarianism can be seen almost as built into a contract of government.

The notion of a *minimum commitment* is an important element in the rationale of utilitarianism, and, if I am right, it particularly applies at the public level. Utilitarianism does in certain respects live up to this promise, in the sense that certainly it rests its judgements on a strictly secular and unmysterious basis, and derives (or at least hopes to derive) its substantial input from what people as a matter of fact want, taking its citizenry as it finds them. But those virtues (to the extent that they are virtues) it in any case shares with certain other systems, as we shall see, which lack some of utilitarianism's characteristic defects. Again, utilitarianism has an appeal because it is, at least in its direct forms, a one-principle system which offers one of the simplest and most

[1] I speak of *government* throughout, as a convenient shorthand for agents or bodies making decisions in the public area. The distinctions between public and private themselves are not meant to be more than extremely rough.

powerful methods possible for eliciting *a* result: its commitment in this regard can also be seen as minimal, in that it makes least demand on ancillary principles. It does, however make enormous demands on supposed empirical information, about peoples' preferences, and that information is not only largely unavailable, but shrouded in conceptual difficulty; but that is seen in the light of a technical or practical difficulty, and utilitarianism appeals to a frame of mind in which technical difficulty, even insuperable technical difficulty, is preferable to moral unclarity, no doubt because it is less alarming. (That frame of mind is in fact deeply foolish; it is even, one might suggest, not very sensible from a utilitarian point of view, but agreement to that may lead once more to the slide in the transcendental direction which we intercepted in the last section.)

The appeal at the social level of utilitarianism's minimal commitments is therefore to some extent not peculiar to it, and to some extent illusory. It is also to some extent real, in the sense that utilitarianism really does make do with fewer ancillary principles and moral notions, but then as critics have repeatedly pointed out, and we shall shortly see, the lightness of its burden in this respect to a great extent merely shows how little of the world's moral luggage it is prepared to pick up. A system of social decision which is indifferent to issues of justice or equity certainly has less to worry about than one that is not indifferent to those considerations. But that type of minimal commitment is not enticing. The desirability of a system of social choice can be considered only relative to what it can reasonably be asked to do, and the simplicity of utilitarianism in this respect is no virtue if it fails to do what can be reasonably required of government, as for instance to consider issues of equity. Certainly the simplicity that utilitarianism can acquire from neglecting these demands is not itself an argument for saying that the demands should not be made.

These are questions that I shall come back to. For the moment, we can note the point that a society disposed to make utilitarian choices in personal morality is very likely to favour utilitarian decision by government, for if they see merit in the first they are likely to see the merit written larger in the second. What about the other way round? The prospect of a society which is utilitarian in government but less so in personal morality is a more recognizable one, and one which lies in a direction favoured by many utilitarian writers. Sometimes it is not easy to tell whether such social arrangements are envisaged by these writers, because a haze hangs over the spot from which the utilitarian assessments are being made, and one cannot see whether the transcendental standpoint has been adopted, and developments in society are being assessed from an imaginary point outside it, or whether, alternatively, a position of utilitarian judgement and decision *within* society is being supposed. Smart's discussion of the utilitarians in a magic society[1] is revealing: they can view society and indeed have an effect on it, but they do not belong to it, and for the best outcome they let the local practices continue. It is not surprising that one should be reminded of colonial administrators, running a system of indirect rule.

If we insist on being told from what actual social spot the utilitarian judgements are being made, and if we form some definite picture of utilitarian decision being located in government, while the populace to a significant extent is non-utilitarian in outlook, then it must surely be that government in that society is very importantly manipulative. For either the government is unresponsive to non-utilitarian demands made on it, and must sustain itself by means other than responsiveness to public demands; or alternatively it has nothing to respond to, because the public's non-utilitarian preferences are directed entirely to private objects. If

[1] Smart, p. 50. Cf. p. 123.

that is conceivable at all, without the public turning out in fact to be utilitarians with non-utilitarian recreations, it will be so only because the government encourages or makes it to be so. In both these cases, the social reality will appear very differently to the utilitarian élite from the way it appears to the ruled. This situation is inherently manipulative, and would very probably demand institutions of coercion or severe political restriction to sustain itself. This is a social and institutional manifestation of that lack of openness which I have already remarked in Smart's proposals.[1] And that is where it has to be written out when utilitarianism returns from the transcendental standpoint to being a political force in society. It is not the ideal observer we have to reckon with, but the unideal agent.

It is worth noticing that the idea of a utilitarian élite involves to a *special* degree the elements of manipulation. It is possible in general for there to be unequal or hierarchical societies which nevertheless allow for respect and decent human relations, so long as people are unconscious that things could be otherwise; but which, once such consciousness has arisen, must inevitably become a different and more oppressive thing.[2] To what extent there are societies genuinely naïve in that sense, is an empirical question, but certainly there could be. But the idea of a society which was ruled by a utilitarian élite and which was naïve in that sense is an absurdity. For utilitarianism is erected on the idea of purposive social action and the alteration of attitudes, by methods and to degrees which only empirical investigation will reveal; and no society whose rulers' outlook was built on that idea could also contain quite innocently the assumption, shared by all, that a division between a utilitarian élite and a non-utilitarian mass was a fact of nature. Individual

[1] See section 6, pp. 123–4.

[2] I have said something about this possibility in 'The idea of equality', reprinted in *Problems of the Self*: see p. 238.

utilitarian theorists may manage to be naïve enough innocently to sustain something like that assumption, but no society could.

I turn now to utilitarian principles of social choice. This is a very large and technical subject, central questions in which are at the heart of welfare economics. I shall not try to enter into these questions.[1] My aim will be merely to produce a rough map of some of the most important issues, constructed on the principle of a journey away from utilitarianism. Starting with the full classical apparatus of utilitarianism, a range of doubts and criticisms can move one through a series of stages until one ends with something which is very little like utilitarianism. An important point about this lies in the fact that there are several stages. I shall group them, for the present purpose, very crudely into three steps. The first is the step from utilitarianism to the recognition that even using what are, in a very general sense, utilitarian-type comparisons of utilities, social decision functions which are not utilitarian are equally possible. This is an important step, since some of the appeal of utilitarianism to those who want definite social results rests on the false assumption (not shared by economists) that utilitarianism is unique in eliciting a decision from data of this kind. The second step casts doubt on the adequacy of utilities, perceived satisfactions and expressed preferences as a total basis for social decision, and entertains conceptions of welfare or happiness which raise more pervasive and less definite problems about interpersonal comparison and aggregation. At the third stage, finally, doubt may break out about the whole enterprise of having, except for very specific and limited purposes, such an ambitious and totalistic social decision machinery in any case; but that is an issue which I shall reach without pursuing.

[1] For a most lucid and helpful account of these matters see Amartya K. Sen's brilliant book, *Collective Choice and Social Welfare* (Holden-Day, San Francisco, 1970).

I start with a formulation of Sen's:[1] "In using individual welfare functions for collective choice, there are at least three separate (but interdependent) problems, viz. (a) measurability of individual welfare, (b) interpersonal comparability of individual welfare, and (c) the form of a function which will specify a social preference relation given individual welfare functions and the comparability assumptions." With regard to (a), one issue, which Smart[2] has mentioned, is whether a cardinal or merely an ordinal measure can be imposed; but it is worth noticing that there is no simple relation between the answers to (a) and to (b), since it is not only possible to achieve some forms of interpersonal comparison with purely ordinal preferences,[3] but also it is possible to have cardinal measures of individual preference which do not yield interpersonal comparisons, but which nevertheless admit of solutions to question (c): this is so in Nash's bargaining model.[4]

Classical utilitarianism makes very strong assumptions with regard to (a) and (b), demanding cardinality in reply to (a) and straightforward interpersonal comparisons in reply to (b); it then offers a simple solution to (c), in the form of maximizing either gross aggregate utility, or else average utility, in the simple sense of the aggregate utility divided by the number of individuals.[5] Now it is possible to run versions of utilitarianism on assumptions less strong than these, and though they might lack classical utilitarianism's celebrated ability to yield, in principle, a definite answer for

[1] *Collective Choice and Social Welfare*, p. 118.

[2] Smart, p. 38.

[3] On this, see Sen chs. 7 and 7*, and also Richard C. Jeffrey, 'On interpersonal utility theory', *Journal of Philosophy* 68 (1971) 647–57.

[4] Cf Sen, chs. 8 and 8*.

[5] Smart seems to hesitate between these importantly different alternatives: p. 28. See also Rawls, *A Theory of Justice* (Harvard University Press, Cambridge, Mass., 1972), pp. 162 ff. I pursue one aspect of the 'average' solution below.

all cases, they might win in other respects, as for instance by being rather less unrealistic; while other systems within this general framework, give different answers to (c) which may convey other advantages. As Sen has put it[1] "Such a general framework . . . does lack the sure-fire effectiveness of classical utilitarianism, which is one of its very special cases, but it also avoids the cocksure character of utilitarianism, as well as its unrestrained arbitrariness".

I am not concerned here with different bases on which utilitarianism, or some version of it, might be run, nor yet with the details of alternative systems, but merely to draw attention to the existence of alternative systems which, while they themselves pay various prices, can do better than utilitarianism in matters on which it is notoriously weak, above all that of equity. Clearly Rawls's maximin principle – regarded here as a principle for comparing social states, rather than for comparing sets of institutions, which is what he offers it as – satisfies this second condition better than utilitarianism does, though it may give implausible results elsewhere; and more generally, the kind of *lexicographic* ordering which Rawls and others have employed – by which some criteria for preference can be brought into play only after others have been satisfied – is more realistic and sophisticated than utilitarianism's gross insistence on summing everything.

In this light, utilitarianism does emerge as absurdly primitive, and it is much too late in the day to be told that questions of equitable or inequitable distribution do not matter because utilitarianism has no satisfactory way of making them matter. On the criterion of maximizing average utility, there is nothing to choose between any two states of society which involve the same number of people sharing in the same aggregate amount of utility, even if in one of them it is relatively evenly distributed, while in the other a very

[1] *Collective Choice and Social Welfare*, p. 104.

small number have a very great deal of it; and it is just silly to say that in fact there is nothing to choose here. It is not a question, it is perhaps worth insisting, of those who insist on a relevant difference here bringing forward a value, while the utilitarian answer involves no values; utilitarian social decisions involve values as much as any do. Nor can we say that such situations will not arise, because for instance inequity will give rise to discontent, which thus reduces the total and average utility. For the objection to an inequitable state is not contingent on the worse-off persons being discontented; on the contrary, their being worse-off provides a ground for their being discontented, and it is a startlingly complacent and conservative conclusion that it must actually be better if, things being inequitable, people are not discontented.

A moralizing argument in favour of maximizing average utility might be this.[1] The moral point of view is impersonal, and abstracts from one's own personal interests, to look at a situation in a universal spirit. But this comes to the same thing as the requirement that in choosing between social states it makes no difference who in particular one is;[2] and this might be represented as the idea that the social state is best in which a citizen selected at random is best off; and this might be thought equivalent to the requirement that average utility be maximized. It seems in any case extremely doubtful that the consequences of impersonality can be represented just in terms of the utility enjoyed by a randomly selected citizen. But even if it could, the argument is invalid

[1] For considerations in this area, see Rawls, *A Theory of Justice*, pp. 164 ff., though the argument offered here is different from his.

[2] Something of this kind may possibly underlie Smart's flirtation (p. 37) with the idea that under moral impersonality, *X*'s sacrificing *X*'s interests should be seen just as a special case of *X*'s sacrificing *Y*'s interests. Why that result is absurd, and hence why impersonality, if it leads to it, is absurd, are questions closely related to the issues of integrity I have discussed earlier.

as a support to the principle of merely maximizing average utility. For clearly there can be two states of society with population and aggregate utility equal in both, but where the probability of picking at random a citizen whose utility falls below the average is much greater in one than in the other; this will obviously be so for a state in which there is a great segregation of utility to a few persons, since in that case there are many more persons with below average utility than in a state in which distribution is more equitable. The argument gains any plausibility it has from another, and different, application of the principle of insufficient reason: it relies on the fact that out of the indefinitely many social states which display a given average utility, the greater number must be states in which the majority of citizens do not differ from each other in utility by too much. But if that fact supports anything in this area, it can support not the principle of merely maximizing average utility, but that of maximizing it granted that differentials are not too great, i.e. it concedes the case for considering distributive issues.

The next step on the journey away from utilitarianism moves us from issues of how one handles utilities and preference schedules, to the question of whether utilities and preference schedules can possibly be all that we are concerned with, even under the heading of individual welfare. We may pass over, though we should not forget, the gigantic difficulty of discovering even ordinal preferences over even private and homogeneous goods. The present difficulties start from the facts that the goods may not be homogeneous, and they may not be private. The principle of the substitutability of satisfactions is basic to utility calculations; it turns up, for instance, and very evidently, in the Hicks–Kaldor compensation test, to the effect that a change is an unequivocal improvement if its beneficiary is made so much better off by it that he could compensate the loser from it and still have something over. It can hardly be an objection to econ-

omics, as economics, that it is about money. But once such principles are seen as *the* principles of social decision, one should face the fact that goods are not necessarily inter-substitutable and consider the case, for instance, of an intransigent landowner who, when his avenue of limes is to be destroyed for the motorway, asks for 1p compensation, since nothing can be compensation. That there must be something which constitutes compensation for a finite loss is just a dogma, one which is more familiar in the traditional version to the effect that every man has his price.

The question arises, again, what objects of preference can be handled by the formulae of social decision. This seems to me a very difficult question, on which not enough is yet known; thus it is far from clear whether games theory can make good its promise to be able to handle any set of preferences, including altruistic ones, without destroying its theoretical basis. We have already met, in section 5, the question of what projects utilitarianism can satisfactorily contain without either collapsing into the evidently restricted and egoistic assumptions of classical Benthamism, or else falling into incoherence about the relations between a man's own projects and the project of utilitarianism itself. In the social field, this same problem emerges once more in the form, particularly, of the question, what degree of social or public content can be allowed to preferences if they are to be straightforwardly part of the input of the social decision function. Groups can hold views about what the state should be like and similar matters of principle or deep concern which they cannot coherently regard as material for a trade-off with other advantages. If they are powerful or determined enough, it is well-known that they can exercise a blocking effect; and structural situations of this kind can lead, for instance, to federal solutions. Now an administrator can view these persons in a utilitarian light, as an obstacle which it costs an indefinitely large amount to remove; but

they cannot regard themselves in that light, and certainly one cannot restrict the notion of 'political thought' to the planning which does regard them in that light – their own thought can itself be political thought. So if utilitarianism is to provide the criterion of rational political thought, it follows that no one should, ideally, think as such persons do. That is to say, utilitarianism once more legislates not just to the handling, but to the content and seriousness, of the projects in society.

As we found in the individual case, so in political decision, utilitarianism is forced to regard 'commitments' (as I previously called them) externally, as a fanatical deviation from the kind of preference which can be co-operatively traded off against conflicting preferences. That might seem in any case a gratuitous evaluation, and an impermissible limitation on the supposed topic-neutrality of utilitarianism's view of preferences. But it might be yet worse. For it might turn out, as I have already mentioned, in discussing the individual case, that the happiness of many men – by criteria of happiness which utilitarianism would itself have to recognize – lay in their identification with these commitments, these self-transcending social objectives which do not allow of trade-offs.

Perhaps humanity is not yet domesticated enough to confine itself to preferences which utilitarianism can handle without contradiction. If so, perhaps utilitarianism should lope off from an unprepared mankind to deal with problems it finds more tractable – such as that presented by Smart[1] in a memorably Beckett-like image, of a world which consists only of a solitary deluded sadist.

There is a different radical problem which arises even if we look at preferences of a more immediately domestic character. However elusive the ordinal structure of an individual's preferences is admitted by utilitarians to be, it

[1] Smart, p. 25.

will naturally be taken to refer to what he does now *actually* prefer. Even if this were ascertainable (and ascertainable without interference, which is a further point), it would fall short of an adequate basis for social decision in many cases, because it might not coincide with what the individual would prefer if he were more fully informed, and if he had some more concrete sense of what things would be like if his preference, or various alternatives to it, came off. Considerations of this kind are often rejected as élitist or authoritarian, and the generous employment of notions of a 'real will' by political manipulators certainly provides grounds for a healthy respect for that kind of objection. But nevertheless, and far short of its more contentious deployments, the point has power. For anyone who admits the role of expert consideration in government – and utilitarians are certainly the last to reject it – thereby admits that an uninformed preference may well fail to coincide with what that same individual would prefer if he became informed. Nor can we accept the idea that it is *just* a matter of people's having established desires, and being informed or not about particular outcomes as realizations of those desires. What one wants, or is capable of wanting, is itself a function of numerous social forces, and importantly rests on a sense of what is possible. Many a potential desire fails to become an express preference because the thought is absent that it would ever be possible to achieve it.

None of this provides an alternative formula for arriving at social decisions, nor could it; but it points to a glib illusion which utilitarianism trades on, and which renders utilitarianism irresponsible – the illusion that preferences are already given, that the role of the social decision process is just to *follow* them. There is no such thing as just following. To engage in those processes which utilitarianism regards as just 'following' is – by a style of argument which, ironically, utilitarianism is particularly fond of – itself doing something:

it is choosing to endorse those preferences, or some set of them, which lie on the surface, as determined by such things as what people at a given moment regard as possible – something which in its turn is affected by the activities of government.

In this, we have a special case of something which is very important. A well-known argument of utilitarianism against criticisms of this kind is that we can agree that everything is imperfect – only roughly discovering preferences and aggregating them, supposing that actual and present preferences are the only relevant preferences, giving strongest emphasis to those preferences which we are theoretically in the best position to handle, treating non-substitutable goods as substitutable, and so on: but that, all the same, half a loaf is better than no bread, and it is better to do what we can with what we can, rather than relapse into unquantifiable intuition and unsystematic decision. This argument contains an illusion. For to exercise utilitarian methods on things which at least seem to respond to them is not merely to provide a benefit in some areas which one cannot provide in all. It is, at least very often, to provide those things with prestige, to give them an unjustifiably large role in the decision, and to dismiss to a greater distance those things which do not respond to the same methods. Just as in the natural sciences, scientific questions get asked in those areas where experimental techniques exist for answering them, so in the very different matter of political and social decision weight will be put on those considerations which respected intellectual techniques can seem, or at least promise, to handle. To regard this as a matter of half a loaf, is to presuppose both that the selective application of those techniques to some elements in the situation does not in itself bias the result, and also that to take in a wider set of considerations will necessarily, in the long run, be a matter of more of the same; and often both those presuppositions are false.

At this point we reach the edge of such large questions as: to what extent should political thought be seen as a matter of systematic principles at all? How far can the application of such principles determine more than very abstract models which the urgencies and complexities of actual political life will make irrelevant? What intellectual structures, such as those of lexicographic arrangement, could be applied to such principles? Are important political changes discontinuous in ways which no one authority acting in an administrative spirit could allow for? In what ways can government, and public control over government, responsibly handle the facts that people's preferences are in some part a function of their expectations, and their expectations in some part a function of what government does? These are real questions, not rhetorical ones, and they are some of the more important, though not necessarily the newest, questions of political philosophy. The relevant point here is that on virtually none of them has utilitarianism anything interesting to say at all; they are questions which start after it has run out.

Utilitarianism is in more than one way an important subject; at least I hope it is, or these words, and this book, will have been wasted. One important feature of it, which I have tried to bring out, is the number of dimensions in which it runs against the complexities of moral thought: in some part because of its consequentialism, in some part because of its view of happiness, and so forth. A common element in utilitarianism's showing in all these respects, I think, is its great simple-mindedness. This not at all the same thing as lack of intellectual sophistication: utilitarianism, both in theory and practice, is alarmingly good at combining technical complexity with simple-mindedness. Nor is it the same as simple-heartedness, which it is at least possible (with something of an effort and in private connexions) to regard as a virtue. Simple-mindedness consists in having too few thoughts and feelings to match the world as it really is. In

private life and the field of personal morality it is often possible to survive in that state – indeed, the very statement of the problem for that case is over-simple, since the question of what moral demands life makes is not independent of what one's morality demands of it. But the demands of political reality and the complexities of political thought are obstinately what they are, and in face of them the simple-mindedness of utilitarianism disqualifies it totally.

The important issues that utilitarianism raises should be discussed in contexts more rewarding than that of utilitarianism itself. The day cannot be too far off in which we hear no more of it.