Week 1 reading:

4

utilitarianism

This chapter deals with *utilitarianism*. We will look at what utilitarianism is, and investigate some of the criticisms that are often made of it, and look at the way utilitarian theory can develop to accommodate such criticism. The emphasis is on the way in which considerations that naturally arise in thinking about morality can lead us to the utilitarian tradition, and how it might then seem worthwhile to get into the business of refining and developing the theory to overcome objections. This distinctively theoretical approach – aiming for a systematic answer to moral questions that is not vulnerable to objections – is not just an "ivory tower" enterprise, but rather a necessary part of fully understanding the role of morality in our lives.

WHAT UTILITARIANISM IS

CASE STUDY: THE UTILITARIAN METHOD

According to the utilitarian approach to ethics, the right action is the one that has the best consequences: specifically, it is that action, out of all available alternatives, that creates the greatest balance of happiness over unhappiness. So in order to work out what the right action is in any situation we need to work out which are the available alternatives (say, A, B, C and D); then we need to work out for each of these alternatives the costs of taking that option in terms of unhappiness caused (say 20 units), and the benefits in terms of happiness caused (40 units); and then we need to work out the net balance of happiness over unhappiness (40–20=20). The option with the greatest balance of happiness over unhappiness is the right action. Now this sounds reasonably straightforward (this basic method is preserved in what is called "cost-benefit analysis", which is often used as an essential part of rational planning in a whole range of organisations). However, it is not quite that simple. For a start, we have simplified things by assuming only four available options. In reality the range of available action-options open to us at any point may be infinite – how do we go about narrowing the range in order to make a reasonable comparison? Secondly, we have simplified by assuming that there are units of happiness and unhappiness – but can happiness be quantified? Thirdly, if we are trying to work out the right action prospectively, that is, in advance of performing it, it is likely that we won't be entirely sure what the consequences of each course of action will be. We will therefore have to work, not with actual values for happiness/ unhappiness, but probabilities. (Probabilities, and the way in which they figure in rational decision-making under uncertainty, are themselves a slightly mysterious quantity – but even if there are such things as objective probabilities, prospectively at least we can only work on an estimate of what the probability is of some consequence occurring.) However, we need some numerical value in order to be able to make the comparison. Thus one way of doing it would be to deal in expected utility, which is basically the unit of happiness that could be brought about multiplied by the probability that it will come about. The utilitarian might recommend (though we will see grounds for questioning this below) that an agent should take the option with the greatest expected utility.

One of the most direct accounts of what utilitarianism involves is given by J. J. C. Smart: for the utilitarian, "the only reason for performing action *A* rather than an alternative action *B* is that doing *A* will make mankind (or, perhaps, all sentient beings) happier than will doing *B*."[1] Let's just comment on a number of aspects of Smart's account. First of all, it makes clear the utilitarian's commitment to *outcomes for happiness*. To put this in more technical language, we can say that utilitarianism is a *consequentialist* theory. For consequentialists, the only things that have value are states of affairs. Consequentialists deny the deontologist's claim that some actions have inherent moral value – as required or forbidden, etc.

To see what's at issue here, have a think about how we might make sense of an act as being forbidden. We could perhaps make clear sense of this idea if we thought that the acts were forbidden by a god. But we may want to make sense of morality in the absence of God – for instance if we are motivated by *naturalism* (the idea that basically what there is in the world is what the natural sciences tell us is in the world). Or we may be unsatisfied with the reasoning behind the claim that acts are wrong or forbidden simply because God has ruled against them (for instance, if one thinks that the reason God rules against them is that they are already wrong).[2] Then we have to find some other way of explaining what is meant by "forbidden." Utilitarians think that it is "spooky" to talk about acts being inherently required or forbidden – it would be to invoke something like a *taboo*. Many societies have *taboos*, of course, and invent stories about how such acts come to be forbidden, but the utilitarian can argue that *taboos* are the kind of thing that, with greater knowledge, we can come to see as merely creations of culture. A *taboo* cannot be valid in its own right. However, the utilitarian does not think that all morality is merely conventional and *taboo*-like. Even if we were sceptical about all our social *taboos*, there remains something real, namely states of affairs involving happiness and misery. Suffering is real even if "thou shalt not" is not. Hence the motivation for consequentialism. Even if we "saw through" all claims about acts being forbidden and required, we should not doubt that some states of affairs are better than others, since some states of affairs plainly involve greater suffering and less welfare than others. Consequentialists hold that it is only states of affairs that have value because they find claims about states of affairs comprehensible while claims about the inherent value of acts are mysterious. For the consequentialist, therefore, if an act has value as right or wrong, etc., then it can only be derivatively, because of the good or bad states of affairs that it produces.

A consequentialist theory is not complete without a specification of which states of affairs are valuable. Utilitarianism tells us that it is the happiness or well-being of sentient beings that is the valuable thing. Consequentialist theories don't have to be utilitarian. You could have a non-utilitarian form of consequentialism that held that what makes states of affairs valuable is freedom or biodiversity or creativity (and where these things are not just valuable because they lead to happiness). Nevertheless, one advantage of utilitarianism is its apparently ready compatibility with naturalism: that we can understand what is good about happiness and bad about suffering, without appealing to anything mysterious or intrinsically valuable. It is part of the basic psychological make-up of sentient beings that they are repelled by pain and attracted by pleasure.

From Smart's definition we can also see that the concern of utilitarianism is with the interests of humanity as a whole (or perhaps sentient beings in general). An important and attractive aspect of utilitarianism is its commitment to equality and impartiality. The utilitarian looks at the goodness of states of affairs, assuming that it is happiness and nothing more that makes them good, and concludes that the happiness of any one person must be just as valuable as the happiness of any other. This is an idea that we may take for granted in the modern age, but it is worth noting how important utilitarian thinkers have been in helping us to get rid of prejudices according to which the interests of some – by virtue of birth, race, sex, social rank, etc. – are more important than those of others. Utilitarianism at its outset was a radical theory that preached that "each should count for one, and none for more than one." The happiness of any one person is just as valuable as that of any other.

Finally, utilitarianism gives us a clear method for getting answers in moral philosophy. Say I am in a situation in which I have promised to spend the evening with a friend. But then another friend phones to say that she needs someone to help her prepare work for the next day, and that she can't find anyone else who will do it. How do I decide what to do? I have to find some way of weighing up the importance of the promise against the importance of helping my friend. But how do I go about this? This process might look a bit mysterious. How do we measure the importance of these different options in order to compare them? Do we even understand what a good answer to this question would look like? The utilitarian has a clear way of dealing with this. For the utilitarian the way in which we work out the right action in any situation is always the same. One should set out the various possible courses of action open to you, and work out the costs and benefits associated with each. Then calculate for each the balance of benefits over costs. The optimal course of action is the one with the greatest balance of benefits over costs. Utilitarianism therefore makes much of what is involved in working out what to do, a straightforward **empirical** matter of calculating costs and benefits. In practice, of course, the calculation might be rather complex, and involve a lot of uncertainty about what will be gained and lost through different courses of action. But we have a clear idea of what a solution to the problem would look like: it is the same as the solution to the question, "Which course of action will lead to the greatest happiness?" Assuming happiness to be something measurable, this approach means that each moral question has a quantifiable answer.

How is utilitarianism likely to apply in practice? We will have a look at this further on. But when we think about issues such as euthanasia, global poverty, animal welfare, and so on, we encounter themes that might lead us to utilitarianism. Utilitarianism, as we have seen, is a moral theory that is rooted in a concern about suffering and welfare. For the radical utilitarian customary morality can seem to be full of conventional rules that prevent or excuse us from doing as much to benefit the world as it is in our power to do. Thus, for instance, through customary morality we have concerns arising from the supposed sanctity of human life that prevent us from maximising the benefit to humanity as a whole. We have ideas about the supposed value of human life as opposed to animal life – even when the humans in question are severely handicapped and thus less intelligent than some of the animals we use for eating and experimentation. We have a distinction between what is required of us and what is merely "saintly" that allows us to escape the responsibility of sharing our lucky inheritance with those less well off than ourselves. The rational solution, for the utilitarian, can seem to be to do away with the "rule-worship" of customary morality and attend directly to what matters – making the world a better place.

However, from the opposing perspective (that is, the *deontological* rather than the consequentialist perspective), the radical approach that the utilitarian suggests looks highly immoral. The characteristic of the utilitarian approach, as we have seen, is that it takes no acts to be ruled out in advance. In other words, when deciding what to do, we must look at the consequences of our acts, and aim to bring about the best results we can. It would be madness, from this perspective, to hamper ourselves by ruling out certain categories of action in advance – to say that we will never (ever) lie or steal or kill the innocent or torture. Any of these acts might become necessary in some situation if we are to do as much good as it is in our power to do. If we committed ourselves never to performing such actions we would be putting ourselves in a situation in which we were sometimes unable to be as effective as we might be. The utilitarian can see no basis for making such a commitment – since, after all, the point of morality is to make the world a better place. However, to someone of a more deontological persuasion, who does see reason to rule certain acts out in advance (as morally unthinkable, say), the utilitarian approach will be not just radical, but radically unprincipled – a sufficiently important end will justify any means necessary.

We can now summarise some of the benefits of utilitarianism as a moral theory. First of all, it gives us a clear and non-mysterious account of what morality is about: producing states of affairs in which there is happiness and freedom from suffering. It gives us a clear and non-mysterious account of how we work out what to do: calculating costs and benefits. It does not set moral limits to what we can do – it just tells us to maximise the good. And it is rooted in impartiality and equality. All of this adds up to a potentially radical and critical theory that we can bring to bear on our habits of moral thinking and action and our social institutions. Let me give two examples of this now.

• UTILITARIANISM IN PRACTICE: PUNISHING AND PROMISING

The first example concerns the institution of punishment. Michel Foucault gives a dramatic historical example of punishment at the opening of his book *Discipline and Punish*.[3] Damiens, an attempted regicide, is paraded in a cart through the streets of Paris to the place of his execution. There he is gradually dismembered, boiling lead poured on his wounds, until his torso is finally pulled apart by horses. This horrific death is a public spectacle. It took place a little over 250 years ago. Foucault's point in using this example is to jolt us into recognising how dramatically our ideas regarding punishment have altered in a relatively short period. Utilitarians can perhaps take some of the credit for this shift in our perceptions. As we saw above, for utilitarians, each person's interests count, and count equally – even the interests of offenders. By contrast, the way Damiens is punished expresses the view that he is nothing, or even less than nothing. Actions that would normally be regarded as barbaric are inflicted righteously, even savoured and enjoyed by the crowd.

The way Damiens is treated suggests a view on which criminals lose their moral status: things can be done to them that cannot be done to people ordinarily. This is a thought that motivates the view of punishment known as *retributivism*. Of course modern retributivists would also see the treatment of Damiens as barbaric. But retributivists share the perspective of those punishing Damiens to the extent that they think that making wrongdoers suffer is not wrong in the way that making others suffer would be. They might explain this by saying that wrongdoers *deserve* to be punished; wrongdoing changes one's moral status. However, the utilitarian stands against this retributive view. The offender does not mysteriously lose her moral status; her happiness continues to count as much as anyone else's. Therefore punishment cannot be deserved or right in itself; punishment can only be justified, as with anything else, by its consequences for happiness. Now on the face of it punishment is highly problematic from a utilitarian point of view: punishment is the deliberate infliction of suffering on an offender for an offence. If such suffering is justified, utilitarians think, then it can only be because it relieves greater suffering. For instance, punishment might be justified if it prevents crime e.g. through deterrence. But on the other hand punishment might well not be justified – if there are alternatives that would be as effective in preventing crime at lesser cost. Utilitarianism might spur us to think of ways in which we could prevent crime without causing such misery.

Here we see that utilitarianism can provide a critical standpoint by which to evaluate social institutions. The critical standpoint is rooted in something objective and quantifiable, namely, consequences for happiness and misery. However, there are at the same time problems with this radical approach. Utilitarianism can appear a positive force since it holds that one should punish only when some good will come out of it. We should not punish unnecessarily, for the sake of it. However, what makes the punishment of an offender necessary – for instance, that it will have great deterrent effect – will in some circumstances give us just as good a reason to punish an innocent person. Imagine a situation in which a person is universally believed to be guilty (you and the innocent party are the only ones who know he is innocent), and in which he can easily be framed by destroying the evidence of his innocence. Furthermore, the person actually guilty of the offence has died, and is therefore no longer any danger. As a utilitarian police chief, you would be faced with a choice between punishing no one, and therefore achieving no deterrent effect, or punishing this innocent party. You know that, while you should never punish when no good will come of it, you ought to punish if it is necessary to bring about a sufficiently important good effect. There is no danger of the man's innocence coming to light. So what bad consequences could come of punishing the innocent? Therefore it seems that utilitarianism would judge that the right thing to do in these circumstances is to punish the innocent party.

This, it might be charged, is another example of utilitarianism leading to immoral results. The utilitarian may of course bite the bullet and say that it is only a commitment to a mysterious deontology that would lead us to rule out punishing the innocent as absolutely wrong. Like any other act, the utilitarian might say, punishing the innocent is something that, though regrettable for the suffering it brings, might be necessary in some circumstances. And then the utilitarian might stress how rare such circumstances are (normally it will be impossible to guarantee that the truth of the person's innocence will be uncovered). But many will remain unhappy with this: the idea of building general welfare on the sacrifice of the innocent may seem totally unacceptable. Furthermore – and this is a point that we will come back to – if it did come out that our police officers were making decisions on utilitarian grounds, and therefore that a proportion of those we thought guilty might be innocent, would we not start to lose faith in the criminal justice system?

Our second example concerns promising. What is involved in making a promise? When you promise someone that you will do something you are not merely predicting that you will do it, nor are you just saying that you intend to do it. Rather it seems that (again, in some rather mysterious way) you are *committing* or *binding* yourself to doing something, in such a way that you are not free to do otherwise than you have promised until the recipient of your promise frees you from it. Our practice of promising seems to involve the thought that simply by uttering some formulaic words ("I promise", "I give you my word") we have made it the case that we have a duty to comply. Now the radical utilitarian looks at this with raised eyebrows. How can I bind myself to do something on Tuesday by making a promise in this way? To a utilitarian this might sound like an odd ritual – and like any other ritual, she might say, it can achieve nothing real. Therefore the utilitarian will take this talk of bindingness with a pinch of salt. Come Tuesday evening the utilitarian will weigh up the utility of complying with the promise against the utility of breaking it in the usual way: setting out the various courses of action and their costs and benefits, etc. Having made a promise one might have created the expectation that one will comply: frustrated expectations (especially where this means frustrated plans) might cause some suffering, and this might affect the costs of breaking the promise. But the promise itself the utilitarian will regard as insignificant. The utilitarian sees through this social convention. Again we see utilitarianism bringing a radical approach to socially accepted rules.

Again, however, there might be disadvantages to taking this sort of attitude to promising. After all, now it seems that a utilitarian cannot make promises. If I know that you are a utilitarian then I will not expect you to keep your promises. I know that if you get a better offer – a better opportunity to promote utility – then you will take it. In that case I will not rely on you. But now think about the range of social interactions that are based on promising – or, what is really a legalistic form of promising, namely, contracts. I work for a university for a month on the understanding that they will pay me at the end of the month. How can I be sure that the university will pay me? Well, because I have a contract with them, and the contract is binding. But what if the university were run by utilitarians and will only pay me what it owes if it is optimal to do so? Would I still be so willing to do my month's work in advance? Or say I lend a book to a student on the basis that she promises to give me it back the next day. Because she has made the promise and takes herself to be bound to return it, I can rely on her to do it. Now imagine that she is a utilitarian. If I lend her the book then I know that she will only return it if that seems to be the optimal thing to do. But I don't want her to do the optimal thing with my book: I just want to have the book back. So if she is a utilitarian I won't enter into a type of cooperation with her that I would have if I knew she had a more deontological attitude towards her promises. This suggests a problematic conclusion, that utilitarianism destroys trust.

Considerations such as these raise a problem for utilitarianism that we will look at again in the next section. This is that utilitarianism is *self-defeating*. The problem can be put like this. Consider two worlds, one in which there are the sorts of social cooperation that promises and contracts enable us to reliably arrange, and one in which there is no such cooperation. It seems plausible that the first world will be happier than the second, since the range of projects that human beings can successfully pursue is dramatically increased once cooperation is possible. The problem is that, if what we said above about the untrustworthiness of utilitarians is correct then the world in which many or most people are utilitarians might end up being like the second of these worlds rather than the first. But the aim of utilitarianism is to maximise happiness. Therefore utilitarianism is self-defeating.

• SOME FURTHER PROBLEMS – THE HARD LIFE OF A UTILITARIAN

These problems for utilitarianism will eventually lead us to a better, though more complex, understanding of how to be a good utilitarian. But before we get on to the solution, let's raise some further problems. The first one concerns the unwieldy nature of utilitarian thinking. Above I lauded utilitarianism for giving us a clear method for working out answers to moral questions. Yes, the reader might have said at that point, with a hint of sarcasm. All one has to do is set out all the possible courses of action open to you, work out for each option every likely cost associated with it (and perhaps the probability that that cost will be realised) and every likely benefit (and its probability), and then work out the course of action that gives the greatest and most likely balance of benefits over costs. Easy! Though of course in reality, and considering the huge number of possible options that are possible for an agent at more or less any moment, this surely involves a mind-bending amount of calculation. Furthermore, this seems to be another way in which utilitarianism could be called self-defeating. The central value of utilitarianism is to make the world a better place. But if we have utilitarian agents spending their whole time engaging in these monumental calculations then they won't have any time left over actually to bring about happiness or alleviate suffering.

The second problem is whether the utilitarian could engage in anything like friendship and other personal relationships. Consider the range of relationships we have that involve some sort of personal attachment or loyalty. When I have a free weekend I phone Phil up to see if he wants to go out. Why him rather than some other, perhaps more needy person? Because he's my friend. I spend time and money on my children rather than any other children, and I phone my parents because they are my parents. I deal with the problems of students on my course because they are my students. The crucial thing seems to be the relation of being *my* friend/child/parent/spouse, etc., which gives us a special connection or importance to one another. This connection, we tend to take it, means that I owe a certain loyalty or special consideration to these people over others (though what such special consideration involves may vary with the type of relationship that it is). Now think of the utilitarian commitment to equality, impartiality and general welfare. Wouldn't the utilitarian look at such claims about "special connections" and find them just as mysterious as talk of the binding nature of promises, and the forbidden nature of *taboos*? This radical utilitarian attack on personal relationships was memorably put forward by William Godwin. Considering a case in which we can only save one of two otherwise doomed people, but where one is Archbishop Fénelon and the other Fénelon's valet (though the valet is your brother or your father), Godwin asks, "What magic is there in the pronoun 'my' that should justify us in overturning the decisions of impartial truth?"[4] Godwin thinks that we must save Fénelon, since he will have by far the greater effect on general welfare. The personal relationship cannot be regarded as morally relevant.

This makes the utilitarian life look like one of rather terrifying austerity, in which morality demands that we deny deeply rooted feelings of kinship, affection and intimacy. The issue of whether the utilitarian has to give up friendships and other personal relationships can be generalised to a concern that utilitarianism makes morality too demanding on us. This relates to the issues we saw arising over global poverty. Moral common sense tells us that we have duties to others to help in certain extreme cases. But it also tells us that we have a right to engage in projects of our own – "in our spare time" as it were. But the utilitarian looks to have no morally "spare" time. Any time in which you are playing football, learning the violin, learning another language, reading novels, and so on, could presumably have been spent doing something that would have far greater impact on the sum of human happiness. Engaging in such activities while others are in need looks to the utilitarian like mere selfish indulgence. Therefore not just our personal relationships but all personal projects are under threat by the utilitarian expansion of morality to cover all areas of our lives.

Although the claim that we should be prepared to give up our personal loyalties and pleasures for the sake of the general welfare may strike us as overly demanding, there is surely something right about the utilitarian attitude. After all, if one agrees with the utilitarian that no one's happiness is any more important than anyone else's – that no one deserves to be happier than anyone else, or certainly not simply because he or she was born in a richer country – then why should one favour one's own happiness or the happiness of one's loved ones over that of anyone else? How could your loved ones be more special than anyone else? Of course, you might say, they are more special to you. But why should your perspective have any genuine moral relevance? Favouring our own means, in the end, that those who "have" continue to have – and to get more – whereas those who "have not" continue to be deprived of a share they could have had. Why is favouring the interests of others when they happen to be friends or family any different from a case of favouring members of one's own race – a case that we admit to be immoral?

Nevertheless there might again be something self-defeating about Godwin's austere attitude. Again, compare two worlds, one in which people form relationships of friendship and love, and take themselves to be free to develop talents and interests (playing the oboe, reading and writing poetry, etc.) and one in which they do not. Given that these relationships and interests are an important source of happiness in a human life, it seems plausible that the first world is happier than the second world. However, won't a world peopled by Godwins be more like the second world? Again, given that the aim of utilitarianism is to make the world a happier place, this seems self-defeating.

• TOWARDS A SOLUTION: RULE-UTILITARIANISM

Let's sum up the criticisms of utilitarianism that we have raised so far. One is that utilitarianism leads to immoral results. Another, that utilitarianism is self-defeating because it would make it impossible to have social customs (like promising) or relationships that promote general happiness. Another is that utilitarianism is self-defeating because its method of decision-making is too cumbersome to employ – that using it would leave no time for promoting happiness. The utilitarian has an interesting line of response to this set of problems. Let's begin with the third criticism, about the cumbersome method.

First of all, let's make the criticism yet more compelling. The problem with utilitarianism, the critic might say, is that it has an unrealistic view of human capacities, perhaps even of human rationality. The utilitarian thinks that we are basically rational maximisers, that we aim to maximise our own interests and/or the general interest. If we were, then we would have to be good calculators, forever assessing our various options and working out costs and benefits. But, the critic says, this **model** – beloved of economists – is a misleading distortion of how humans actually think and behave. Human beings do not calculate at each step. Rather they follow *patterns* of behaviour, they form habits, and such patterns save us from the impossible task of calculating everything from scratch at every moment. There is some scope for evaluating our habits, of course, but only at relatively rare moments of reflection: often there is no time and it would be counter-productive to be always crippled by the need to reflect. Furthermore, it might be said, the utilitarian overestimates the power of individual rationality to overcome social context. In reality we are social creatures, and the power we have to think beyond our social milieu is limited. Largely the ways we think and act are conditioned, if not determined, by social structures. The patterns of behaviour we form and follow are social patterns: ones we share with others. In the vast majority of our behaviour we follow socially-instituted practices and rules rather than making it up for ourselves according to our own calculation of utility. Nor could we realistically imagine it being any other way.

If true this may seem to be a pretty devastating criticism. But as we will see the utilitarian has a way of turning it to her advantage. Say the utilitarian agrees with all of these empirical facts about the way human beings behave and think. Of course, the utilitarian might say, being a naturalistic theory utilitarianism must work with human beings as they are, and not demand the impossible from them. But if that is the case then these observations about how humans think and behave, if true, just make it clearer what form utilitarianism ought to take. Therefore, the only viable form of utilitarianism is one in which the fact that human beings are rule-followers rather than rational maximisers is built in at the start. The problems that we have raised for utilitarianism have largely come about because we have concentrated on what we can call *act-utilitarianism*, the view that the utilitarian method of working out what is right applies to each action. But now imagine *rule-utilitarianism* that accepts that human beings will follow patterns of behaviour as though following rules, and applies the method to those rules rather than to the individual acts. On rule-utilitarianism we compare the utility of people in a society following different possible rules rather than taking different possible actions. Moral thinking becomes more about the design of a society structured by various (rule-governed) practices and institutions – in which we are choosing those practices, institutions and rules that will produce greater utility when people engage with them – than about the governing of individual conduct. The governing of individual conduct still goes on, of course: that is the point of a moral theory. But it is indirect: through the rules rather than by a straightforward calculation of the utility likely to result in your case. In order to work out what you, as an individual, ought to do in a particular situation, you have to work out what rule it would be optimal for everyone to follow in such situations, and then act according to that rule.

One formal way of stating the difference between act- and rule-utilitarianism is to look at their differing *criteria of right action*. We can state act-utilitarianism thus: an action is the right one in a situation **if and only if** it would result in greater utility than any alternative available action. For rule-utilitarianism, by contrast, an action is right if and only if it *falls under a rule* the *general following* of which would result in greater utility than an alternative available rule. A less formal way of seeing the difference is to look at some of the examples we previously found problematic. For instance, we previously said that utilitarianism seemed to be self-defeating since agents who operate according to act-utilitarian procedures – weighing up the utility of each action – would not be trusted by their fellows, and that a society populated by such agents would be less happy than one in which promises are kept. This is grist to the rule-utilitarian mill. The rule-utilitarian method *starts* by comparing the two worlds, the one in which promises are, and the one in which promises are not, able to be kept; if it turns out that the former is a happier world then "Keep your promises" is a rule that ought to be followed. The same goes for family and friendship relationships. If the world in which such relationships are formed is a happier one than a world in which they are not, then some rule such as "Favour your friends" ought to be followed. Favouring your friends, keeping your promises: these acts can turn out to be right on rule-utilitarian grounds.

Rule-utilitarianism therefore promises to solve a number of the problems associated with simple versions of utilitarianism. It can solve the problem of unmanageable calculation and the other problems that threatened to make utilitarianism self-defeating. Indeed the rule-utilitarian might say that what these cases of apparently self-defeating consequences show is simply that it is not optimal always to follow the rule (which we have assumed to be set out in act-utilitarianism), "Always assess the consequences of your actions and try to act optimally." What our discussion brings out is that the general following of this act-utilitarian rule does not lead to the best available consequences, and that instead following rules that seem to have more in common with deontological commandments and the demands of customary morality can have better results.

Hence another advantage of rule-utilitarianism is that it reduces the appearance that utilitarianism leads us to acts that are immoral. Consider again the case of punishing the innocent. When we argued that utilitarianism would lead to punishment of the innocent whenever doing so would bring about the same benefits (of deterrence say) that would justify the punishment of the guilty, we were assuming that law officials would make their decisions on a case-by-case basis. But now look at things through a rule-utilitarian lens, in which we are assuming that individuals are following the rules of practices or institutions. So now take two worlds: One of these is a world in which there is an institution that gives officials the discretion to punish the innocent when they judge it to be sufficiently advantageous to general welfare (call it "telishment", since the basis for punishment is straightforwardly *teleological*, or consequentialist, rather than *deontological*). And the other is a world in which there is our familiar institution of punishment, in which the rule is that all and only the guilty are to be punished. In which world is the sum of happiness greater? In his paper "Two Concepts of Rules", John Rawls argues that the world of telishment will be the worse,[5] in part because it will involve giving legal officials unaccountable authority to make secret decisions to punish the innocent when they are believed to be guilty – authority that could easily be abused – and in part because it will leave citizens feeling insecure about whether they might be telished and unsure whether to condemn or pity those apprehended by criminal justice. The happier of the two worlds, Rawls thinks, would be the world with the practice of punishing only the guilty. Thus "punish only the guilty" is a rule that we should abide by since it is the rule that, out of the alternatives (or at any rate, the alternatives we have considered), gives the best consequences.

Rule-utilitarianism therefore appears to take the sting out of some of the deontologist's strongest criticisms of "unprincipled" utilitarianism. Rule-utilitarianism gives us a morality that does contain *principles*, and many of the principles it contains are ones to which we are intuitively committed. But while rule-utilitarianism explains and allows us rationally to endorse our commitment to these principles, it does so without invoking anything mysterious like the deontologist's "thou shalt." Even if one shares the consequentialist's view that *taboos* are mysterious when taken at face value, one can still accept a range of principles: on the rule-utilitarian view what justifies these principles is nothing more than the fact that generally following them brings about good states of affairs. You don't have to believe in anything more mysterious than happiness and suffering to explain the authority of principles.

• CRITICISMS OF RULE-UTILITARIANISM

Nevertheless the deontologist won't be happy with the rule-utilitarian solution. For the deontologist there is still the problem that rule-utilitarianism makes the validity of the moral rules too **contingent**, too accidental. Consider for instance a rule-utilitarian justification of some basic rights – to life, property, basic freedoms, etc. The idea of *natural rights* – according to which it is forbidden to treat persons in certain ways by virtue of their metaphysical/moral status – was famously written off by Jeremy Bentham as "nonsense on stilts." Natural rights theory gives a deontological account of rights as a (metaphysically mysterious?) "thou shalt not" attaching to certain beings. But the rule-utilitarian can give a justification of rights without this metaphysical peculiarity. The rule-utilitarian, as we have seen, looks at two worlds, one in which rights are respected and one in which they are not, and then argues that, *because the world in which they are will be happier than that in which they are not*, rights are valid moral rules. In other words, for the rule-utilitarian our practice of respecting rights is instrumental to our producing the happiest outcomes. The deontologist's problem has to do with what follows the "because", as a reason to respect rights. For the deontologist it is problematic that the rule-utilitarian thinks that rights are valid *only if* the social practice of respecting them leads to greater happiness. For the deontologist this gets it all wrong: according to the deontologist, the basis of rights is the dignity or sanctity of the human personality, something that will remain important even if it does not always make for a happier world. Imagine for instance that the world actually would be a happier place if there was an institution of benevolent slavery. Under such slavery some people would be denied some of their basic human rights (though otherwise they might be well looked after). Would there nevertheless be something morally wrong with such a happy world? The deontologist thinks there would: the utilitarian fails to explain why slavery is wrong *in principle* – that it is contrary to the freedom and dignity of the human spirit, say (more on this in the next chapter, which will be on Kant) – even if it has good consequences. Of course the utilitarian thinks that this talk of dignity or sanctity just sounds like a *taboo*, and raises the question of how there could be anything in our material world that, as the deontologist thinks, we are *forbidden* to enslave. The rule-utilitarian would rather avoid this high-flown metaphysical rhetoric, and settle for the empirically more verifiable claim that societies that keep slaves are far more likely to be unhappy.

The rule-utilitarian may think that she has gone some way to trying to accommodate the sources of the concern that utilitarianism leads us to immoral action, all without compromising the basic utilitarian and consequentialist outlook. If there are aspects of the theory that still leave the deontologist unhappy, perhaps we just have to say that these show a profound difference in moral orientation, something that cannot be settled by moral argument. Perhaps more surprisingly, however, rule-utilitarianism might also leave a utilitarian unhappy, and it is this type of criticism that I will go on to look at now.

When a utilitarian looks at rule-utilitarianism, he may feel that it is just not utilitarian enough, that it has sacrificed too much of the radical context-sensitivity of the theory and returned to rule worship. Utilitarianism in its simple form is context-sensitive in the sense that it is always the outcomes available in a particular situation that determine whether an act is right rather than its conformity to a set of predetermined rules. However, it now appears that, with rule-utilitarianism, we do have a theory according to which acts are right insofar as they conform to the rules. The rules themselves may be determined by circumstances and outcomes in the utilitarian manner, but is that really enough? J. J. C. Smart provides an example that, though it may appear contrived, will help us to articulate just what seems unsatisfactory from the utilitarian point of view. One of the advantages of rule-utilitarianism, we said above, is that it can explain why we ought to obey moral rules like "keep your promises" – for utilitarian reasons. But take a situation in which breaking a promise would clearly have better consequences than keeping it (even though of course the practice of making promises is more useful than not having the practice).

First, recall that breaking promises can have all sorts of bad consequences. It can offend the promisee, and cause suffering by the frustration of her plans. It can destroy trust (since if everyone did it no one would believe anyone else when they promised), and this would be bad since promising is a socially beneficial practice: society would be less happy and less well off if we could not rely on one another. And furthermore, since promising is a useful practice, it is useful that we have a strong psychological aversion to – a gut reaction against – breaking promises. If we keep on breaking promises we can weaken our psychological habit of instinctively keeping promises. Smart gives us a case in which these bad consequences are either irrelevant or clearly outweighed by the good consequences.

Imagine you are stuck on a desert island with a man who makes you promise that, if you get off the island and he does not, you will make sure that his vast fortune is given to the local riding club. As it happens the man dies but you do get off the desert island. It occurs to you that you could do much more good if you gave the man's money to a local hospital rather than to the riding club. Should you keep your promise? If you are a rule-utilitarian you will judge that the right action is to keep your promise, since by doing so you will conform to the socially beneficial rule. But consider the pros and cons of the situation. You will not offend the promisee by breaking the promise, since he is dead. You will not weaken the social practice of promising by breaking the promise, since no one need be told that you made the promise in the first place. You may weaken your own instinctive commitment to promising (and to telling the truth, since you will have to lie about making the promise), which is a bad consequence. But surely isn't this a minor evil in comparison with the good you can do by giving the man's fortune to the hospital? Smart therefore thinks that we should not be too hasty to discard act-utilitarianism. He thinks that rules might be useful to the act-utilitarian agent. But they should only be thought of as guides, "rules of thumb", rather than as part of what makes acts right and wrong.

This example may seem contrived. But we could apply the same reasoning to the case of torture in order to bring out the key issue of principle. The rule-utilitarian will endorse a rule against torture, since it is overwhelmingly likely that a world without torture will be happier than a world with it. But now imagine that we are in a situation in which we can only avert the destruction of a large city by getting information out of a terrorist. Can we use torture? There are some deontologists who will say no, that even in this case there are some things we cannot do to our fellow human beings. But it is characteristic of the utilitarian tradition to look to the results. Smart's point is that it seems odd that the rule-utilitarian lines up with the deontologist in judging that torture in this situation is wrong. Rule-utilitarianism is too insulated from the outcomes of particular cases.

Smart thinks that this should lead us to an improved act-utilitarianism rather than rule-utilitarianism. We can explain this alternative utilitarianism in more detail by making a distinction between utilitarianism thought of as a *criterion of right and wrong,* and utilitarianism thought of as a *decision procedure* or a guide to action. In explaining act- and rule-utilitarianism in this chapter we have assumed that, in deciding how to act, the utilitarian agent should aim to follow her theory, taking it at face value. We have assumed, in other words, that alongside telling us what is right and wrong a theory attempts to guide our action, and that these two jobs are one and the same. If the theory says that the right act is the one that produces the best consequences then the good utilitarian should let herself be guided by this theory, and try to produce the best consequences. Because trying to produce the best consequences seems to be a policy that will have *bad* consequences, we assumed that we had to change our account of which acts are right and wrong, so we moved to rule-utilitarianism. However, rule-utilitarianism looks as if in certain situations it might also have bad consequences, from a utilitarian point of view. However, what Smart points out is that none of this ought to affect the basic utilitarian understanding of what makes acts right and wrong – this is just a matter of the consequences they produce. While this question of what the criterion of right and wrong is, is a *moral matter,* the question of how agents ought to think and behave in order to be good utilitarians – i.e. to produce the best consequences – is an *empirical matter,* something that we could find out by trial and error. The discussion above suggests that good utilitarians ought not to behave as act-utilitarians or as rule-utilitarians: neither of these policies is likely to maximise benefits. So we might suggest that while utilitarians ought to be committed to a certain understanding of right and wrong, it might be an open question how they ought to behave, whether and to what extent they ought to follow rules, form friendships, go in for self-improvement, respect rights. Saying that it is an open question does not mean that there is no right or wrong answer, that it is up to each person to make their own decision. It is simply that, as with many empirical matters, we don't yet know what the answer is: it is a matter for further experiment and investigation.

To sum up, we have been looking at the way the utilitarian tradition can help us think about what morality involves. This tradition holds to the central thought that what morality is really about is human welfare, and that moral standards are only important because human welfare is important. It is a tradition that rejects the seemingly mysterious idea that some acts are intrinsically wrong in favour of a kind of naturalism, rooted in our understandable tendency to give importance to human welfare. The utilitarian theory can throw up some odd moral results, so it is important to see that this commitment to naturalism is a central part of the utilitarian's overall justification. Depending on the circumstances, utilitarianism might say that we are justified in torturing, enslaving, lying – things that moral common sense strongly rejects. But faced with criticism the utilitarian can ask what the alternatives are – does moral common sense simply take for granted a realm of "thou shalt nots", belief in which cannot really be defended once we begin to question it? The radical utilitarian can respond to her opponent's moral qualms by simply biting the bullet, accepting the **counter-intuitive** conclusion and seeking to puncture the opponent's assumption that the moral standards they are talking about can be real. Given the guiding belief that consequences for welfare are all that matters, we have looked at the way in which utilitarianism might develop into an attractive and workable theory.

• SOME CONCLUDING THOUGHTS ABOUT THE NATURE OF HAPPINESS

Before we finish we should raise a question about happiness that may have been bothering the reader all the way through. One of the advantages of utilitarianism that I have been stressing throughout the chapter is its naturalism. The utilitarian shows us how we can believe in morality while at the same time believing that the world contains nothing more than what the natural sciences tell us about. A related attractive aspect of utilitarianism is that it makes moral questions answerable in the same way that other empirical questions can be answered: by looking at the way the world works and how the most happiness might be produced. Moral questions, for the utilitarian, are just like other technical questions in which we are trying to find out the most efficient way to further our ends. This might seem a great advance on the position we started out with in this book, in which it appeared that our moral reasoning and the faculties with which we find out about the moral world were mysterious and unlike any other type of reasoning that we have. Or at any rate, this would be a great advance if we could make one assumption: *that happiness can be measured*. If happiness is something real, concrete and quantifiable then the utilitarian's project will seem plausible. But a potential problem looms here. For if on the other hand it turns out that happiness is something indefinable, or something that itself involves morally controversial judgements, then the utilitarian view will not look so naturalistic.

The thing is that there are some things in the world that are clearly suited for empirical investigation: they are "out there", their nature waiting to be discovered. For instance, when we want to know how many tables there are in a room, we look into the room and count: there is an independently definable answer for us to discover. But there are other things that, when we make judgements about them, we already assume an evaluative perspective. For instance, if I judge that the arrangement of tables in the room is "orderly" or "messy" then I am invoking an *evaluative perspective*. The issue for the utilitarian is whether the question "What will bring about the greatest happiness?" is more like the empirical question "How many tables are there in the room?" or more like the evaluative question "Are the tables orderly or messy?" As I have been explaining the motivations of utilitarianism, there is a strong tendency for utilitarians to think that questions about happiness must be basically empirical. After all, one of the strengths of utilitarianism was supposed to be the way that it shows how ethics might be non-mysterious. If it turns out that judgements about happiness themselves involve evaluative judgements then all the problems that we started with are raised again. How do we find out what happiness is? By which faculty? What sorts of facts are facts about happiness if not empirical facts? Hence we find that utilitarians tend to support some view of happiness according to which happiness can be in some way measured.

For instance Bentham thought that happiness simply consisted in feelings of pleasure. It may be hard for us to measure feelings of pleasure, but they are at least a clear part of the empirical world. We do have a rough knowledge of what the causes of pleasure and pain are, and we might reasonably expect that psychologists and so on will become better at understanding them, so we may well become better able to predict which courses of action will lead to the greatest sum of pleasure. Bentham himself offered a hedonic calculus, a comprehensive set of criteria by which the basic idea of maximal overall pleasure should be measured. However, Bentham faced criticism that has to be addressed by any attempt to come up with a quantifiable conception of happiness. The criticism is that when you come up with a view on which happiness is measurable you will be unable to explain why happiness is morally important, or at any rate the only morally important thing. Mill was addressing this criticism when he developed his famous doctrine of "higher" and "lower" pleasures.[6] Bentham had made happiness measurable by equating it with pleasure, but now faced the criticism that he had come up with a "philosophy fit for pigs." The concern was that the perfect world for the Benthamite utilitarian is one in which people become more like pigs, in which they settle for easy pleasures rather than striving for ideals ("push-pin", a child's game, rather than poetry), but that this, like Huxley's *Brave New World,* would be a nightmare rather than a Utopia. Pleasure, the concern was, may be measurable but is not the only thing that we want in a good life. We might even say it is not the only thing that makes us *happy*. In response Mill argued that both parties were right: Bentham is right that happiness is pleasure; but the critics were also right that some ways of life are "higher" than others. Mill interprets the critics' view that some things one can do with one's life are more important than others as a claim about pleasure: these ways of life bring, not just more pleasure, but a better *kind* of pleasure. However, the idea of a better kind of pleasure is one that might seem again to simply assume that we can make evaluative judgements rather than showing how such evaluative judgements are only really ever empirical judgements.

Modern-day utilitarians are less likely to argue that happiness consists in pleasure. The most popular conception of happiness today is preference satisfaction. In other words, the view is that the more you get what you want (satisfy your preferences) the happier you are. Some preferences can be stronger than others, of course, so this contributes to how happy they make you. But the preference-satisfaction view is meant to be an improvement on hedonism since it takes account of the fact that we don't just want pleasure, we also want a range of other things (and not just for the sake of the pleasure they give us). At the same time, people's preferences have a decent claim to be things that we can find out about empirically, either through observing behaviour (on the assumption that people tend to act to satisfy their desires and preferences, and tend to show their priorities in their behaviour) or by asking them what their preferences are. Nevertheless there is still a question about whether someone who gets what they want is necessarily happy, or whether giving people what they want is the only thing that is morally important. Sometimes people want things because they have false beliefs about them, they imagine that getting them will be better than it actually would. If they realised what it would really be like they wouldn't want them any more. Furthermore, isn't it the case that sometimes people want trivial, pointless things that distract them from what is really important in life? Is the world really made a better place by giving people what they want if what they want is worthless? Therefore it is not entirely clear that preference satisfaction is an uncontroversial or merely empirical way of understanding what happiness is. But without an empirical understanding of happiness utilitarians cannot use naturalism as a defence against the counter-intuitive results that their theory sometimes throws up.

• CONCLUSION

Utilitarianism holds that what is important is happiness and the avoidance of misery. Acts are right only as means to this end, and wrong only if they do not bring about as much happiness as they might. Utilitarianism has various advantages as a moral theory. It gives a clear and non-mysterious explanation of what morality is about and how we find out what is right and wrong (i.e. by calculating outcomes). It also gives us a clear critical standard by which to evaluate current social practices and customs. However, utilitarianism has faced various criticisms. Among these we considered: that it recommends immoral actions; that it is time-consuming to apply; that it prevents us from having the sorts of projects and relationships that give meaning to our lives. We looked at rule-utilitarianism as a way of solving these problems. Rule-utilitarianism recommends that we follow socially beneficial rules rather than attempting to assess consequences for ourselves. However, rule-utilitarianism is criticised by deontologists for failing to explain the true basis of moral rules. And it is criticised by utilitarians for not being utilitarian enough. The strongest form of utilitarianism may be that which sees utilitarian moral theory as providing a criterion of right action, and recognising that it is simply a further empirical matter how a utilitarian agent ought to think and act in order to maximise right action. Nevertheless utilitarianism is only plausible if there is some way of measuring happiness. But if we think of happiness as something that is measurable – feeling pleasure or satisfying preferences – it is not clear that that is the only thing that is morally important.

• QUESTIONS FOR DISCUSSION

1 Can you state in your own words the reasons a critic might have for thinking that utilitarianism is self-defeating? Is this a good objection to utilitarianism or is there a version of utilitarianism that escapes it?

2 Would it be wrong to have an institution of slavery if it made society happier overall?

3 Do we ever know what the consequences of our actions will be before we perform them? Even after we have acted, are there not an infinite number of effects of any action? How could we know about them all? Do questions like these raise a serious problem for utilitarianism?

4 In this chapter, utilitarianism has been presented as a naturalistic theory that demystifies the "thou shalt" of deontology. But does it really escape the need for moral "oughts"? For instance, utilitarianism tells us that we ought to promote happiness. And isn't this to go further than merely stating the fact that one state of affairs contains more happiness than another? In Chapter 4 of *Utilitarianism*, Mill argues that happiness is "desirable" because each person desires it, and that, as each person's happiness is good for that person, so the good of all is a good for the aggregate of persons. Do you find these claims plausible? Do they help to show that utilitarianism escapes the need for any mysterious sense of "thou shalt"?

5 What is happiness? Can it be measured? Is happiness so important that its maximisation constitutes the whole of our moral duty, as utilitarianism holds?

• FURTHER READING

J. Bentham, *Introduction to the Principles of Morals and Legislation* (various editions), is a classic statement of utilitarianism, both for its unflinching recognition of some of the problems of the theory, but also for the ingenuity and consistency in addressing them.

Another classic utilitarian text is J. S. Mill, *Utilitarianism* (various editions). Mill takes more seriously the aspirations of "higher" culture and tries to show how they can be combined with utilitarianism.

For a good introductory defence of utilitarianism, see W. Shaw, *Contemporary Ethics: Taking Account of Utilitarianism* (Oxford: Blackwell, 1999).

J. J. C. Smart, "Extreme and Restricted Utilitarianism", reprinted in J. Rachels, *Ethical Theory 2* (Oxford: Oxford University Press, 1998), gives a good explanation of rule-utilitarianism plus some key criticisms of it from a utilitarian direction.

For an honest and resourceful response to the claim that utilitarianism would justify immoralities like slavery if the overall consequences were good enough, see R. M. Hare, "What is Wrong with Slavery", reprinted in P. Singer (ed.), *Applied Ethics* (Oxford: Oxford University Press, 1986).

P. Railton, "Alienation, Consequentialism and the Demands of Morality", reprinted in S. Scheffler (ed.), *Consequentialism and Its Critics* (Oxford: Oxford University

Press, 1988), is a good example of modern "indirect" utilitarianism, which aims to explain and resolve the problem that utilitarianism is incompatible with friendship and love. It also addresses the wider question of the utilitarian agent's relationship to their society.

• NOTES

[1] J. J. C. Smart, "An Outline of a System of Utilitarian Ethics", in J. J. C. Smart and B. Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973), p. 30.

[2] This is often called the *Euthyphro* problem, since it was raised in Plato's dialogue of that name. See the discussion, "The *Euthyphro* Problem", in Chapter 7.

[3] M. Foucault, *Discipline and Punish: The Birth of the Prison*, trans. A. Sheridan (Harmondsworth: Penguin, 1991).

[4] W. Godwin, *Enquiry Concerning Political Justice and Its Influence on Morals and Happiness* (various editions), Vol. 1, bk 2, ch. 2.

[5] J. Rawls, "Two Concepts of Rules", *Philosophical Review* 64 (1955): 3–32.

[6] Here we return to some of the themes we encountered in the section "Higher Pleasures?" in Chapter 1.