9
Utilitarianism

"I don't want to! Why should I?"
"Because more people will be happier if you do than if you don't."
"So what? I don't care about other people."
"You should."
"But why?"
"Because more people will be happier if you do than if you don't."
(Katherine Tait, *My Father Bertrand Russell*, 1970: 184–5)

9.1 Moral theory and deliberative practice again

What place in a good person's deliberations can a moral theory have? Towards the end of Chapter 8, we explored this question as it arises for the moral theory called virtue ethics. We noted that the virtue ethical account of rightness says this:

> *Virtue ethics:* An action is right iff it is the action that a virtuous agent would characteristically do in the circumstances.

(Or something like that. We developed some complications about what the exact formula should be in §8.6, but we need not revisit these now.)

We also noted in Chapter 8 that any plausible moral theory is bound to be about more than simply action. It needs to have something to

say, for instance, about emotion, response, choice and deliberation as well. Hence we can produce formulas like the following that parallel the account of rightness:

> An emotion is right iff it is the emotion that a virtuous agent would characteristically feel in the circumstances;
>
> A response is right iff it is the response that a virtuous agent would characteristically produce in the circumstances;
>
> A choice is right iff it is the choice that a virtuous agent would characteristically make in the circumstances;
>
> A deliberation is right iff it is a deliberation that a virtuous agent would characteristically perform in the circumstances;
>
> and so on.

This last formulation raises again the interesting question for virtue ethics that I first raised in §8.3: will a virtuous agent's deliberation ever include materials such as the virtue ethical account of rightness itself? In other words, will the theory itself be explicit in the deliberative life of the sort of agents that the theory says there ought to be?

This relation between moral theory and actual deliberative practice seems to be a very important one for normative ethics. The reason why is because any moral theory needs to have a stable relation to deliberative practice. That is, if we cannot explain how a moral theory is supposed to help us deliberate, or if the contents of the moral theory are contradictory to the contents of deliberative practice, then it is hard to see a use for the theory.

Of course, we do not have to suppose that the moral theory is involved in deliberation directly: it might be involved only indirectly, by being part of how we reflect on our deliberations "in a cool hour", as I suggested was possible for virtue ethics. But if it cannot even be involved in that sort of reflection, then the moral theory's usefulness is questionable, even if it is true.

These questions about a moral theory's relation with actual deliberative practice are important ones for every moral theory. One thing that will emerge from this chapter is the importance of these questions for

utilitarianism. But before we get to that, let us begin with a review of what utilitarianism says.

9.2 Utilitarianism in outline

Utilitarianism has a basic form, and then many variations designed to deal with objections to the basic form. One way of stating the basic form of utilitarianism has already been offered in §8.1:

> *Utilitarianism*: An action is right iff it promotes the greatest happiness of the greatest number.

A somewhat fuller statement of the basic form of utilitarianism sees it as the conjunction of four separable theses:

> Utilitarianism = maximalism + welfarism +
> aggregationism + consequentialism
>
> where the component theses are defined as follows:
>
> *Maximalism*: It is always obligatory to take the best option available.
> *Welfarism*: What is good is happiness/welfare.
> *Aggregationism*: We can measure and quantify happiness/welfare, both within lives and between lives.
> *Consequentialism*: The goodness of any option depends only on the goodness of its consequences.

Here maximalism tells us that we may never aim at less than the best. Welfarism and aggregationism together tell us what "the best" is, and how to measure it. In Mill's famous phrase, "it is the greatest happiness of the greatest number", which we find by comparing the amounts of happiness/welfare that various alternative possible states of affairs will contain for various numbers of people. And consequentialism tell us that "the best" is a state of affairs, a set of consequences, that is achievable by action.

(A note on the word "consequentialism": there are other ways of using the word besides the one laid down by my definition. The term is

sometimes no more than a rough equivalent to "utilitarianism". It can also mean, as I think Brad Hooker intends it to mean in *Ideal Code, Real World* [2003], "a position family-related to utilitarianism, only more plausible". In Anscombe's original sense – she invented the term in "Modern Moral Philosophy" [Geach & Gormally 2005: 184] – the word means something else again, namely, the denial that there is any morally significant distinction between actually intended and merely foreseen consequences. This variety of usage need not be confusing, provided the reader is aware of it.)

All four of utilitarianism's component theses have such a strong and obvious intuitive appeal that, to many moral theorists, utilitarianism seems like "the only game in town". It can seem like the only moral theory worth taking seriously, or the default moral theory, the starting-point from which all departures need to be justified by argument:

- *Maximalism*: If you *can* do the best, it is hard to see why you would settle for less.
- *Welfarism*: Happiness can hardly be a *bad* thing; if moral theories do not contribute to human happiness, what is the point of them?
- *Aggregationism*: We may not be able to measure happiness scientifically, or use instruments to detect and compare happiness levels in different people. But we do generally know when one person is happier than another, how much of one sort of happiness is a fair exchange for how much of another sort, and so on. And we must know this kind of thing, not just in order to run utilitarianism, but in order to be able to say what distributions are fair.
- *Consequentialism*: It would obviously be mad not to consider the consequences of what you propose to do. And why, more generally, would anyone do anything *except* to bring about this or that consequence?

Something like the theory that we get by combining these theses can be found in many philosophical texts, including the writings of Francis Hutcheson and Adam Smith in the eighteenth century, and Spinoza's *Ethics* in the seventeenth century. Long before these writers, something similar is already visible in Socrates' proposal in *Protagoras*:

> Like a man who is skilled in measuring weights, put pleasures in the scale together with pains, bearing in mind their distance or

> proximity; then say which side weighs more. If you are weighing pleasures against other pleasures, choose the larger and weightier ones; if you are weighing pains against other pains, choose the lesser and the smaller ones. If you are weighing pleasures against pains, then the right choice depends on whether or not the pains are outweighed by some pleasures which follow on them. Are near pains outweighed by distant pleasures, or distant pains by near pleasures? If so, the action to be done is the one that displays those features. But if, in the case of some action, near or distant pleasures are outweighed by near or distant pains, then don't do that action. (356b–c, my translation in Chappell 1996: 98–9)

Here Socrates is suggesting alterations to the virtue-based view of ethics that came naturally in his society towards a view that will, supposedly, make it easier to decide between alternatives. The theory that he sketches is one in which we treat all goods and all evils as pleasures and pains. This gives us a moral "currency" in which everything can be priced against everything else (everything is *commensurable*). Socrates is thus rejecting a view that comes naturally to a virtue ethicist: *value pluralism*, the view that there are many different goods and no way of making them commensurate. Socrates here rejects value pluralism in favour of something like aggregationism. Socrates also, apparently, collapses the distinction, important in virtue ethics, between means–end and constitutive relations (§5.1): he suggests that every action is a way of bringing about a state of pleasure or pain (or both). This moves him away from the variety of different sorts of possible action that a virtue ethicist will want to recognize, towards the more uniform view of action as instrumentality that I call consequentialism. Finally, on Socrates' proposal (which may or may not be Plato's proposal, or a proposal that Socrates is offering seriously, or one that he offers elsewhere) there would always be something irrational about preferring a lesser sum of good to a greater one. There is nothing in virtue ethics to oblige us to take this view, which is very close to what I call maximalism. (Socrates' suggestion is a view about rationality, and maximalism is a view about moral obligation; but the two views clearly go naturally together.)

Utilitarianism, then, not only is a very intuitively appealing moral theory but also has a long history of attracting the adherence, or at least the interest, of some very eminent philosophers. Its four constituent

theses raise all sorts of interesting questions. In the next four sections I examine them in turn.

9.3 Welfarism

Utilitarianism's historical development out of something very like virtue ethics is clear from the passage I have just quoted from Plato's *Protagoras*. It is also evident from the typical (if not universal) utilitarian commitment to a naturalistic theory of the good, in some ways rather like – and in others rather unlike – the one I described in §8.2 as a key feature of most versions of virtue ethics. Utilitarians too typically answer the question "What are we really talking about in ethical discourse?", which we saw in Chapter 7, by a form of naturalism: as I have called it, welfaristic naturalism. This is the view that the real content of moral discourse is given by claims about what contributes to human welfare or happiness or pleasure. As to what human welfare, happiness or pleasure might be, more on that shortly. But notice that there is no necessary commitment in utilitarianism to anything like the biological naturalist account of these notions that is usually found in virtue ethics.

Apparently this naturalism about the good is both a strength and a weakness of utilitarianism, for the same sort of reasons as it is both a strength and a weakness of virtue ethics. On the one hand, it helps utilitarianism to avoid the meta-ethical problems that are supposed to result from recognizing a non-natural, "special moral reality". On the other hand, it makes utilitarianism prone to the usual objection to ethical naturalism: that if people just happened to want something different from what they actually want, or to take pleasure in something different, then we would have to say that *that* was the good, possibly with very counter-intuitive consequences. (What if everyone started enjoying seeing other people in pain?)

In parallel with the point about virtue ethics that we noted at the end of §8.4, it is technically possible for a utilitarian to avoid these characteristic problems of ethical naturalism, simply by dropping the naturalism. Utilitarianism can be, and often is, presented simply as a theory about *what everybody wants*, no matter whether anyone takes that to be good or not. This form of utilitarianism – *preference*

utilitarianism – has seemed very attractive to many moderns, since it does not commit us on the difficult issues of meta-ethics. This way, we can have a form of utilitarianism that is consistent with any degree of meta-ethical subjectivism or relativism.

However, the account of the good that the classical utilitarians offered was an objectivist and a naturalist one. Besides facing the usual questions for objectivism and naturalism, that account faces specific questions of its own. Let us look to Mill to clarify these questions.

Mill takes "pleasure and freedom from pain" and "happiness" to be equivalent terms, and both names for the human good:

> The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure and the absence of pain; by unhappiness, pain and the privation of pleasure. ([1863] 1962: 257)

Indeed, in *Utilitarianism* Mill offers what he calls a proof that happiness is the good. According to Mill, happiness is proved to be the good by the fact that it is what we desire; indeed, it is the only thing that we *can* desire. In Mill's own words:

> The only proof capable of being given that an object is visible, is that people actually see it. The only proof that a sound is audible, is that people hear it … In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it … no reason can be given why the general happiness is desirable, except that each person … desires his own happiness. [Here] we have not only all the proof which the case admits of, but all which it is possible to require, that happiness is a good. (*Ibid.*: 289)

Note that last indefinite article. This passage takes us only as far as the first step of Mill's argument: that happiness, that is, pleasure and freedom from pain, is *a* good. This on its own is not very controversial; almost anyone will agree to it. The view that happiness is *a* good is not the utilitarian view. Utilitarianism involves the different and stronger

view that happiness is *the* good: as Mill himself puts it, that "happiness is desirable, and the only thing desirable, as an end; all other things being only desirable as means to that end" (*ibid.*).

So Mill needs to do more than this to prove utilitarianism. In his own words, he needs "to show, not only that people desire happiness, but that they never desire anything else" (*ibid.*). To do this, Mill broadens his notion of happiness. It started out (see the extract above) as meaning the same as "pleasure and freedom from pain". It now seems to mean something much less specific than that:

> The ingredients of happiness are very various, and each of them is desirable in itself, and not merely when considered as swelling an aggregate. The principle of utility does not mean that any given pleasure, as music, for instance, or any given exemption from pain, as for example health, is to be looked upon as means to a collective something termed happiness, and to be desired on that account. They are desired and desirable in and for themselves; besides being means, they are part of the end.
>
> ([1863] 1962: 289–90)

We can find happiness, Mill is arguing, in all sorts of ways; all sorts of things are instances or constituents of happiness (cf. the distinction we saw in §5.1 between means–end and constitutive relations). Desiring any of these things – music, or health, or whatever – is desiring an instance of happiness. Therefore, it is desiring happiness.

We might wonder whether this proves that happiness (in the sense that Mill gives the word) is the only thing we ever desire or can desire. Why could someone not choose to be miserable? In fact, do people not choose misery all the time? Or alternatively, why could someone not choose something that is not opposite to happiness, like misery, but simply different from it: for instance, might someone not want to be the person who discovers the proof of Fermat's Last Theorem, without particularly *caring* whether he was happy or unhappy? "If one has one's *why* in life, one can put up with almost any *how*", commented Nietzsche; "man does not strive after happiness, only the Englishman does that" ([1889] 1968: 12).

This point that someone could systematically choose things other than pleasure or happiness – including, at the limit, pain and unhap-

piness – is not addressed very explicitly by Mill. There are two things that we might say on his behalf. One is to claim that people who choose things other than pleasure or happiness can still be choosing in a way that makes sense from a welfarist point of view. It makes sense because the welfarist can simply accept the *paradox of hedonism* – the familiar fact that, often, the best way to find happiness, or pleasure, is not to look for it or focus on it, but to go after something else instead. Your happiness or pleasure will then be all the greater, because you are not fixating on happiness or pleasure. This might look plausible of someone who is absorbed in the routine of his sailing, not in the pleasure that sailing gives him, but finds his pleasure in sailing precisely in this absorption. But even here, there is a doubt about whether someone who sails for this sort of reason is truly described as aiming at pleasure, rather than at sailing. And that doubt is magnified in the case of someone who explicitly says, as the mathematician bent on proving Fermat's Last Theorem or the self-destructive person might, that he "doesn't care about his own happiness".

The other thing we might say is this. Mill's real point is not that it is *impossible* for people to choose to be miserable or unhealthy, or to choose to sacrifice a good deal of possible happiness in pursuit of a mathematical proof. His real point is that it is *unintelligible*. Action, on Mill's view, is only intelligible in so far as it is directed at some instance or other of happiness. If we want to explain to people why they should aim at music or health or proving a theorem, not at being depressed or ill, the only possible way of doing it is to point out to them that music and health and proving a theorem will all, in their different ways, make them happy, whereas being depressed or ill will not make them happy. Thus all explanation of our reasons for doing anything refers back, ultimately, to happiness; actions that cannot be explained in this way cannot be explained at all.

If this is Mill's real point, then one standard criticism of him, deriving from G. E. Moore, seems mistaken. Moore's criticism fastens on Mill's (admittedly rather loose) use of the word "desirable" in the sentence quoted above: "the sole evidence it is possible to produce that anything is desirable, is that people do actually desire it". Moore claims to detect an equivocation in this: Mill, he says, mixes up "desirable" in the sense of "*capable of* being desired" with "desirable" in the sense of "*such as ought to* be desired", and hence commits a version of what is

sometimes termed the is–ought confusion (§§6.2, 6.6). (Or, in Moore's own terminology, Mill commits the "naturalistic fallacy"; see Moore [1903] 2002: ch. 1. Moore is one of the chief originators of our modern difficulties in being quite sure what we mean by "natural" and "naturalism", which is one reason why I do not use his terminology.)

Moore thinks that Mill's mistake is to infer from a premise about what *is* desired to a conclusion about what *ought to be* desired. But if I have reconstructed Mill's argument accurately, this seems to be a misguided criticism, because Mill does not make this mistake. What Mill does is move from the premise that all rationally intelligible action is aimed at happiness to the conclusion that happiness is the good. If we add the plausible premise that the objective of all rationally intelligible action is the good, then we can credit Mill with a valid and persuasive argument:

1. All rationally intelligible action is aimed at happiness.
2. Whatever all rationally intelligible action is aimed at, is the good.
3. Therefore, happiness is the good.

So Mill seems to escape Moore's criticism. But perhaps he still faces a different problem about the notion of happiness. We saw above how Mill broadens his conception of happiness to include pretty well everything that anyone could pursue: "The ingredients of happiness are very various", he tells us, and "They are desired and desirable in and for themselves". (Notice Mill's phrasing here, by the way. Does the fact that he can write "desired and desirable" show him conflating these notions, as Moore charges, or distinguishing them?)

The difficulty about this broadening, a critic might say, is that it puts Mill in danger of emptying the notion of happiness of any particular content. When happiness or welfare is *this* broadly conceived, it just becomes another word for "whatever people aim at". So saying "All rationally intelligible action is aimed at happiness" just means "All rationally intelligible action is aimed at whatever it is aimed at". But that is an empty claim. There is no more content to it than there would be to saying "All cars are going wherever they are going". And so, the charge would be, Mill's argument becomes empty.

To this the utilitarian can reply that it is not empty to say that "All rationally intelligible action is aimed at whatever it is aimed at". Another

way of putting that is to say "All rationally intelligible action has some aim", and this claim is clearly not an empty one. (Any more than it would be empty to say "All cars are going somewhere", which is more or less the same claim as "All cars are going wherever they are going".) We might agree with the critic that Mill has widened the notion of happiness so far that virtually anything can go into it, except that there is still one important constraint on what can be an ingredient of happiness. This, of course, is the requirement that actions that are part of happiness have to have *rationally intelligible* aims. Happiness has become whatever we can *reasonably* and *comprehensibly* aim at.

This is not such a broad conception of happiness that there is no content to it at all. But surely it is a much broader conception than Mill started with. After all, on his original conception of happiness, "happiness" was a near-synonym with "pleasure". And "pleasure" surely cannot mean all the things that "happiness" now seems to mean for Mill.

Actually it can, if you choose to use the word that way. Like "happiness", "pleasure" is a deeply ambiguous word. At one end of its range of meanings, "pleasure" can mean a specific physical sensation. At the other end, it can mean something much less specific, such as "enjoyment" or, even less specifically, "what you find worthwhile". ("Why are you doing 100 press-ups?" "For pleasure." This reply can make sense even if the person doing the press-ups is in a lot of pain!) It would obviously not be a very promising theory that claimed that everything we do is done as a means to pleasure in the physical-sensation sense. But a very different theory, and a much more promising one, will result if we claim instead that everything we do is done either as a means to something that we enjoy or find worthwhile, or as an instance of something that we enjoy or find worthwhile (cf. §4.3).

Most of the time, "pleasure and the avoidance of pain" in the physical-sensation sense seems not to be the objective of our actions. But that does not stop pleasure in some broader sense from being our objective. When I do the crossword, I am not aiming at "pleasure or the avoidance of pain" in the sense that I am when I take an aspirin or ask someone to scratch my back; I am aiming at completing the crossword. All the same, if you ask me whether for me doing crosswords is "business or pleasure", my reply will be "pleasure". And if you ask me why I do crosswords, I may well reply "because I enjoy them". I do not mean by this that doing crosswords is a *means* to enjoyment for me; I mean, rather, that doing

crosswords is an *instance* of enjoyment for me. Nonetheless, when we explain my reasons for action, pleasure (in the sense of enjoyment) is firmly in the picture.

So in fact "pleasure" is not necessarily a different concept from "happiness". Interestingly, both concepts are ambiguous in the same sort of way. At one end of the spectrum both words mean one particular kind of feeling, but at the other end, they both mean something much more general and much vaguer: perhaps just that an activity is done *with enjoyment*. However, neither term is completely empty of meaning, even at this general or (as it is sometimes called) "adverbial" end of their respective spectrums of meaning. So Mill can be justified in equating happiness with pleasure, and in claiming that happiness and/or pleasure is the point of all action, provided it is the general senses of the two words that he has in mind.

The same goes for Mill's predecessor Jeremy Bentham, who defines his technical term "utility" like this:

> By utility is meant the property in any object, whereby it tends to produce benefit, advantage, pleasure, or happiness (*all this in the present case comes to the same thing*) or (*what comes again to the same thing*) to prevent the happening of mischief, pain, evil, or unhappiness. ([1789] 1962: 34, emphasis added)

Mill and Bentham are perfectly entitled to speak of the objective of action as "happiness", or "pleasure", or "utility", or "welfare" in this broad and inclusive sense. For as we have seen, all that is meant by "happiness" (or the other terms) in this broad sense is "whatever makes any action worthwhile or rationally intelligible". So long as we think that actions *can* be worthwhile or rationally intelligible, there is no need for anyone to dispute this.

9.4 Aggregationism

What might be questioned is Mill's and Bentham's claim that happiness in this broad sense can be aggregated. Aggregationism, recall, is the distinctively utilitarian claim that we can measure and quantify happiness or welfare, both within lives and between lives. It is important to

utilitarianism, as normally understood, that this claim should be true, because utilitarianism is a *maximalist* view: it tells us to do the best. But we cannot do the best unless we can work out what the best *is*; and working out what the best is will usually involve us in measuring and comparing amounts of pleasure or happiness or welfare both within and between different lives.

How are we to do this measuring and comparing? If "pleasure" or "happiness" means a sensation or a feeling, it might not be too hard to measure this feeling. We do not yet have a scientific instrument that can scan the brain and detect physical pleasure levels, or other sensations, within. But such technology is quite possibly not too far away from being invented. And if and when it comes along, a scientist will be able to *tell* us when one person is in a state of greater or lesser pleasure than another person, or than himself at some other time.

The difficulty is that, as we saw in §9.3, this is not what Mill and Bentham mean by "pleasure" or "happiness". What they mean – or are best understood as meaning – by "pleasure" is "whatever makes any action worthwhile or rationally intelligible". But this does not look like something that can be measured with brain-scanning scientific instruments! Worthwhileness/rational intelligibility is much too diverse for that. There are so many different things that can make an action worthwhile or rationally intelligible that it is very hard to see how we might get them all on the same scale of measurement.

The utilitarian has a response to this difficulty. It is to suggest that the measurements and comparisons of different degrees of happiness or pleasure are not supplied by scientific instruments. Instead, they are supplied by *us. We* compare possible sources of pleasure or happiness, and say what we find gives us more pleasure than what, and by how much. The rankings that maximalism needs are supplied by general agreement.

One obvious question about this suggestion is: what if people do not agree about how to rank pleasures or forms of happiness?

The utilitarian can go in one of two ways here. One way is for him to say "If people don't agree, then there is no answer". This is to allow the possibility that the rankings of degrees of happiness or pleasure have gaps or indeterminacies in them. By looking to see what people agree on, we can know that two pleasures *A* and *B* are both pleasanter than any of the pleasures *X*, *Y* and *Z*, and both less pleasant than any of the pleasures *P*, *Q* and *R*; but we cannot know which of *A* or *B* is

more pleasant. The result of taking this option is an aggregationism that allows indeterminacies. It is also something like a pure subjective preference-utilitarianism: a view that bases its values solely on what people actually prefer, whether or not they are in some sense right to prefer it. Bentham seems close to this sort of position when he famously claims that "prejudices apart, the game of push-pin is of equal value with the arts and sciences of music and of poetry" (1825: ch.1).

The other way is for the utilitarian to say "If people don't agree, then we need to find an expert who can determine the answer". The utilitarian who goes this way is likely to claim that there are *no* gaps or indeterminacies in the pleasure rankings. Or at any rate he is likely to claim that there are fewer gaps than you might expect, and that where there really is a gap in the rankings this is not just because people disagree about how to rank two pleasures, but because even an expert ranker cannot find a reason for ranking them one way or the other. This gives us a utilitarianism that is more objectivist, and less purely preference based (and more elitist). On this view, it is possible to be *wrong* in my preferences, because the expert judge of pleasures provides a criterion of right and wrong preference. Mill seems close to this sort of position when he famously argues, against Bentham, for a distinction between "higher" and "lower" pleasures. To adjudicate between these sorts of pleasure, Mill appeals to the expertise of those who have experienced both sorts, who (he tells us) will report that even a small amount of higher pleasure is better than a large quantity of lower pleasure:

> It is better to be a human being dissatisfied than a pig satisfied; it is better to be Socrates dissatisfied than a fool satisfied. And if the fool, or the pig, are of a different opinion, it is because they only know their side of the question. The other party to the comparison knows both sides. ([1863] 1962: 260)

A further question about aggregationism fastens on the difference between *intra*personal and *inter*personal aggregation. Maybe (someone might say) we can make some sense of the idea of an individual person's deciding which of various possible actions he would most like to do, on the grounds that it would give him more pleasure or happiness than the alternatives. But is it equally easy to make sense of the idea of deciding between whole groups of actions that are going to affect

whole groups of people? As we might put it, each of us can know in his own case how much a given pleasure means to him; but he cannot be so sure how much that same pleasure will mean to someone else. So even if we can aggregate happiness within one life, does it follow that we can aggregate happiness across lives? We might think that there was something faintly totalitarian about even trying. Surely, we might say, it is up to each individual to decide *for himself* what makes his own life go best. It is, after all, *his life*.

That it is harder to aggregate across lives than within them is widely acknowledged among utilitarians. That it is impossible to aggregate interpersonally is a much stronger claim, and a much more controversial one. As with the last question, there are two ways a utilitarian can go to explain how interpersonal aggregation might be possible. One is to imagine us all voting together on what arrangement we think will make most of us most happy, and say that it is the verdict of the majority that settles how best to aggregate interpersonally. This explanation of interpersonal aggregation tends towards a subjectivist and preference-based view. The other is to imagine an expert whose role it is to *work out* what arrangement best captures the most happiness for the most of us. This time, Bentham seems to be on the side of the expert (indeed he writes in a way that suggests that the utilitarian is himself the expert):

> Sum up all the values of all the pleasures on the one side, and those of all the pains on the other. The balance, if it be on the side of pleasure, will give the *good* tendency of the act [or practice] on the whole … if on the side of pain, the *bad* tendency of it on the whole. ([1789] 1962: 40)

This latter explanation tends towards an objectivist (and more elitist) view of interpersonal aggregation. Both options have their adherents in contemporary philosophy.

A third option that is also interesting is to allow *intrapersonal* aggregation while rejecting *interpersonal* aggregation: to say that we can measure and compare amounts of happiness within lives, but not between them. If we take this line, we might end up with a theory that is utilitarian within lives, but non-utilitarian between lives. That is an interesting possibility, but it is not really a form of utilitarianism at all. (You could take this route and end up with a theory that was [roughly]

Kantian or contractarian overall, but had a utilitarian-looking fragment within it, namely the bit of the theory that deals with individuals' choices within their own lives.) Since our topic here is utilitarianism, I shall say no more about this possibility here.

One final question about aggregationism is this: why should I *care* about the interpersonal aggregate of happiness? Utilitarianism tells me (as we shall see) that what I have reason to do is bring about the greatest possible amount of happiness in general. But *why* do I have reason to do that? Why do I have reason to do more than bring about *my own* happiness?

This question is the utilitarian version of the well-known problem of egoism, a problem closely related to Chapter 3's "Why be moral?" problem. In another way, the question brings us back to the issue of internal and external reasons raised in Chapter 6. One way of responding to it is to look for an argument – as I did in §6.7 – for thinking that everyone has the same internal reasons, and that these include reasons to bring about the general happiness. Another way of responding to it is to accept that we do not necessarily have internal reasons to bring about the general happiness, but argue that we do have *external* reason to do this, and that it would be nice if our internal motivations were to line up to match this external reason. This seems to be the strategy of argument that is pursued by, for instance, Peter Singer (1995), who takes "the moral point of view", defined in the utilitarian way in which he understands it, to be something that we may or may not want to adopt.

I have no space here to do more than note these possibilities for a utilitarian theorist who is developing his aggregationism. The main point for the moment is simply that, in one of these ways or another, the utilitarian does need to be able to make aggregationism workable. For unless he can measure the good, he will not be able to deploy the third component of his view: his maximalism. I turn to maximalism in the next section.

9.5 Maximalism

Maximalism, recall, is the thesis that it is always obligatory to take the best option available. I said in §9.2 that some intuitions tell in favour of this thesis: if you can go for the best, why go for anything less? On

the other hand, some other intuitions tell against it. It is natural to think that there are some actions that are "above and beyond the call of duty": it is good if we do them, but we are not wrong not to do them. Heroic or superhuman virtue cannot reasonably be *demanded* of people, as maximalism suggests it should be. (But perhaps the maximalist can separate his commitment to the view that "It is always wrong not to bring about the best" from his views about blame. Perhaps he can say that blame too is only justified if it brings about the best, and that blaming people for not bringing about the best is not, itself, a practice of blaming that brings about the best. So, at any rate, many utilitarians have suggested.)

The maximalist thesis that it is always obligatory to take the best option available presupposes that there always *is* a best option. This presupposition, however, can be challenged. One sort of challenge arises from the sort of difficulties for aggregationism noted in §9.4. If the goodness of alternative options is always a matter of the happiness or pleasure that they produce, but there is no good way of ranking instances of happiness or pleasure, then there will be no best option either.

The maximalist can point out, in reply, that no one would deny that we can meaningfully compare *some* options for goodness or badness, and that very often their goodness or badness has something to do with happiness or unhappiness, pleasure or pain. It is better to hand out chocolate bars than hand grenades in the playground, and at least part of the reason why is clearly that chocolate bars cause children pleasure, whereas hand grenades cause them pain. The maximalist can say that his position is simply a generalization of these moves.

This generalization faces another kind of problem, about the difficulty of knowing that any option you identify *is* the best one. There seem to be indefinitely many ways of doing anything, and all of them can be counted as different options. Since they are different options, the question arises which of them is the best. This commits the agent who takes maximalism seriously to working out which of them is the best. But that looks like a task of indefinite length.

Of course, when faced with this task a maximalist can respond "Enough deliberating: we've surely fixed on the best option by now". But that you have found the best option is a substantial assumption. Unless you are quite sure that it is correct, maximalism gives you no permission to act on it. It always seems possible that a little more

thought would have enabled you to see a better option than the one you have so far identified as best; if so, you should have deliberated further, and taken that option. There again, it always seems possible that a little more thought would have been a waste of time. The trouble is that it is very hard to tell which of these alternatives is true. And if we cannot tell, then we cannot answer the question "Should I act on this option, or deliberate further about what to do?": not, at least, if we are determined to take *literally* the best option available.

Here a maximalist might remind us of the distinction between *deliberative procedure* and *criterion of rightness* that I made in §8.5. Surely, he will say, we need not suppose that agents must actually, in deliberation, think through all the options in order to determine which of them is best. Certainly agents could not function if they even tried to do that. But so what? Maximalism says that the agent acts rightly iff he takes the best option. It does not commit the agent to deliberating in such a way as to sort through all the options until he has spotted the best one. My criticism, he will say, confuses deliberative procedure and criterion of rightness. Keep them distinct, and we can accept the maximalist criterion of rightness without worrying too much whether anyone can follow that criterion of rightness in practice, or indeed whether anyone ever does, strictly speaking, act in a way that the maximalist will count as correct.

This response raises two further questions. First, if the maximalist agrees that we should not attempt to identify best options in real-time deliberation, then how, according to him, *should* we deliberate? His most promising answer seems to be: "By whatever means of deliberation maximizes good consequences". But this answer prompts another question: which method of deliberation maximizes good consequences? And that question, apparently, is merely a new variant of the original problem. Instead of choosing an action from an indefinitely large range of alternatives, we are now trying to choose a method of deliberation from an indefinitely large range of alternatives. Has the maximalist escaped the problem by modularizing his theory? Or just relocated it?

The second further question is this. If the maximalist's criterion of rightness is not to be used in actual deliberation, then when *is* it used? (Cf. a question I asked about virtue ethics in §8.5.) The difficulty about identifying the best option does not go away when we turn from deliberative processes to a criterion of rightness. If the best option is so

hard to identify, then not only is it hard to identify in real-time deliberation; it is equally hard to identify in calm reflection in a quiet hour.

"It is not to be expected", writes Bentham, "that this process [of evaluating consequences] should be strictly pursued previously to every moral judgement, or to every legislative or judicial operation. It may, however, be always kept in view" ([1789] 1962: 66). What, we might wonder, does "keeping in view" mean here?

Apparently Bentham's idea in distinguishing deliberative procedure from criterion of rightness like this is to suggest that there is some possible standpoint from which we can look over our deliberative practices and our choices in general, to try to make sure that they are optimal even when we assess them from a detached viewpoint. But if it is difficult to identify the best option, it must be equally difficult to identify, or know we are occupying, this detached viewpoint.

Here a point that I made in §9.1 comes back into focus, about the crucial importance, for ethical theory, of actual deliberative processes. The present difficulty for maximalism is that it proposes a criterion of rightness ("Do the best option") that, by definition, is likely to come adrift from any actual deliberative process. In practice, maximalists tend to get round this difficulty by identifying the best option *that there is* with the best option *from whatever small and manageable set of options they actually work with*. However, the best option *that maximalists can think of* might easily not coincide with the best option *that there is.* After all, the options that people actually come up with, when they try to generate a list, are likely to be shaped by their antecedent view of the world – their prejudices, their obsessions, their blind spots and so on – in just the kind of modularizing way that Table 2 (§8.3) describes, only without the prior, and separate, attempt to become virtuous. The danger here for maximalism is the danger of seeming to deliberate more rationally than they really do.

These difficulties about maximalism might prompt the question: can a utilitarian reject it? The answer is yes, in principle. There is obviously nothing incoherent about a moral theory that accepts welfarism, aggregationism and consequentialism, but denies maximalism. The main difficulty in rejecting maximalism is simply what will replace it.

Maximalism, remember, is an answer to the question "Which options should we take?". If a utilitarian accepts no other theory to fill the gap left by maximalism, then the utilitarian will have a hole at the heart of

his moral theory. If he accepts some other theory, then there will be an interesting question about what theory this will be, and how stable its relation will be to the rest of his view.

It is easy enough to propose a *satisficing* version of utilitarianism, like this:

> *Satisficing utilitarianism:* An action is right iff it is any action that produces good enough consequences.

The difficulty with satisficing utilitarianism is to say what counts as "good enough", and why. The utilitarian has two options in answering this question: he can argue that there is some uniquely correct answer to this question, or he can just stipulate an answer. Either way, a maximalist utilitarian is likely to be his fiercest critic. As before, the intuitively gripping question is: why should we ever settle for less than the best we can do?

So much for maximalism, for now. I shall have more to say in §9.7 about the relation between moral theory and actual deliberative procedures. These comments will be relevant to the assessment of maximalism, although they will arise from the assessment of utilitarianism overall. In §9.6 I turn to the fourth and last component thesis of utilitarianism: consequentialism.

9.6 Consequentialism

Consequentialism, I said in §9.2, is the thesis that *the goodness of any option depends only on the goodness of its consequences*. Note that "only". Consequentialism goes much further than the obvious and intuitive view that the consequences of what you do are morally important. It goes further, even, than the plausible view that the consequences are *always* important. Consequentialism is the view that *only* the consequences are *ever* important.

One question for consequentialism is: how are we to identify the consequences of any action? Not all actions *have* consequences, not, at least, in the same way. Some actions are like pushing a button that operates an ejector-seat. They are straightforward cause-and-effect sequences, with a means–end structure. The point of the action of

pushing the button is the consequence of pushing the button, the end to which pushing that button is a means: namely, the operation of the ejector-seat. But other actions do not have this sort of structure. Playing the violin may be a way of enjoying yourself, something you do "for pleasure". But that does not mean that violin-playing is to enjoyment as pressing the button is to the flight of the ejector-seat. Violin-playing is not a means to the end of enjoyment; rather, it is an instance of enjoyment. (Here again is the distinction between instrumental and constitutive relations that we saw in §5.1.)

Perhaps the consequentialist can accommodate this by a small technical wiggle. Can he not say, truly enough, that the consequence of my violin-playing is that it will then be true that I have played the violin? However, this does not seem quite right. My reason for playing the violin is not that I want to *have played* the violin, but that I want *to play* the violin. So perhaps what the consequentialist ought to say is that my violin-playing is, so to speak, *its own* consequence. And likewise with other actions that, like violin-playing and unlike pressing the ejector button, do not aim at anything beyond themselves.

Perhaps this idea can be made to work. Whether or not it can, the consequentialist still has another question to face up to. The normal notion of a consequence – before we stretch the notion in the theory-driven way I have just suggested – is the notion of a future effect of what I do now. So a consequence-based reason for action has to be a reason for action that comes from the future. But, someone might say, our reasons for action do not *have* to come from the future; they can come from the past or the present as well. For example, training now for a marathon next month is action on a reason that comes from the future. My reason for training is that, when the marathon comes, I want to be physically ready for it. But greeting you warmly as you arrive in my room is action on a reason that arises in the present: here you are, and I want to express my friendly feelings towards you. And punishing or thanking someone for what they did yesterday is action on a reason that comes from the past: you did whatever-it-was, and it is now time to respond appropriately to your deed. Consequentialism finds it natural to assume that all reasons for action come from the future. But there seem to be reasons that are best understood as coming from the past or the present. The question remains: how should consequentialism handle these reasons?

A related criticism is this: the consequentialist thinks not only that reasons come from the future, but also that they come from future *states of affairs*. But thinking again about the case of violin-playing, we might wonder whether it is any state of affairs at all that I am aiming at in playing the violin. Isn't my aim an *activity*, not a state of affairs? And similarly, if I greet you in a friendly fashion, might we not say that my aim is a *relationship* (with you), not a state of affairs? Or if I thank you or punish you, isn't my aim an *expression of my views*, not a state of affairs? When you think about it, consequentialism seems to invest quite a lot in the idea that value-bearing entities can only be states of affairs, not items from other categories such as, for instance, activities or relationships or expressions. But this idea has been questioned, for example by Williams: "We do not merely want the world to contain certain states of affairs … it is a deep error of consequentialism to believe that this is all we want" (1985: 56).

A different sort of question about consequentialism is raised by the point that we need to decide *which* consequences we are going to look at. In the definition of consequentialism that I have given, there is no justification for doing anything but looking at literally the total consequences of any option that is taken. But if we do that, we will face two very difficult questions. One is the question we have already considered: what counts as a consequence of something? And the other is: when can we be sure that we have seen the *last* consequence of anything? Is my writing these words a consequence of the Norman Conquest, and if so, is it something that the Normans should have taken into account in deciding whether to invade England in 1066? (It is not exactly obvious that William the Conqueror and his army cared about any moral considerations, let alone those relating to 900 years in the future.) Again, a deed might have nothing but mildly good consequences for 1000 years, and then catastrophic consequences for 10,000: the discovery of petroleum, for instance. It is a very familiar complaint about consequentialism that it is hard to know how to assess future consequences.

To deal with this problem, consequentialists can distinguish between *subjective* and *objective* rightness. The subjectively right thing to do is the one based on the calculus of foreseen consequences; the objectively right thing is the one based on the calculus of all consequences, seen and unforeseen. We are always justified in doing the subjectively right

thing, provided we do our best to ensure that "subjectively best" and "objectively best" coincide.

One question worth considering about this manoeuvre is this: how do we draw a stable line between foreseen and unforeseen consequences? Notice that the more I reflect and try to find out, the more I will foresee. So I will face a series of decisions about whether it will produce better consequences for me to deliberate further in order to get a clearer idea of the consequences of my various options, or to assess them as I understand them now. These decisions look as if they might be pretty tricky.

Another way of dealing with the problems of consequentialism brings us to consider an important subgroup of theories within the utilitarian group: namely the "indirect" or "rule" versions of utilitarianism.

9.7 Rule utilitarianism

The basic consequentialist thesis says that the rightness or wrongness of any action depends only on the goodness of its consequences. But consider the group of contemporary theorists whom we may call rule utilitarians. (This is my name for them, in line with my definitions. They usually call themselves "rule consequentialists", because they use "consequentialist" in the broad sense of "utilitarian-style but not actually utilitarian").

Instead of looking at individual actions, rule utilitarians look at *policies* of action. They propose that the rightness or wrongness of any general policy of action depends on the goodness of the consequences of generally implementing that policy. This revision helps the rule utilitarian to explain why we should keep various sorts of basic moral rule, such as "Do not steal", even in situations where a more act-oriented utilitarian would have to say that we ought to steal, because the one-off action of stealing will be better for overall utility.

If we adopt this sort of view, we can produce a version of utilitarianism on which the account of right action will read something like this:

> *Rule utilitarianism:* An action is right iff it is an action in accordance with the set of rules that, if they were accepted as a general policy, would produce the best consequences.

This formulation suffers from the problem noted above: the problem of ever being sure that the consequences of anything are literally the best possible. Indeed rule utilitarianism suffers from a more severe form of that problem than "straight" utilitarianism (act utilitarianism). For it is bound to be harder, and more complicated, to compare the possible consequences of following a whole variety of alternative general policies or codes of ethics, than to compare the possible consequences of a variety of alternative single actions.

Rule utilitarians could deal with this problem by softening the maximalism inherent in their account of rightness to produce a satisficing theory parallel to the one displayed in §9.5:

> *Satisficing rule utilitarianism:* An action is right iff it is an action in accordance with any set of rules that, if they were accepted as a general policy, would produce good enough consequences.

Satisficing rule utilitarianism still faces the same problems about knowing what consequences any action or policy will produce. There is also the problem that I noted with the previous satisficing theory, about knowing how good "good enough" is. Nonetheless, this theory might seem less threatened than ordinary rule utilitarianism by difficulties about identifying the best of all possible general policies.

Another way of raising problems for rule utilitarianism is to probe the phrase "general policy"; that is, to ask questions about the different possible degrees of compliance with any rule. Any set of rules is likely to produce one set of consequences if everyone complies with it, but quite another set of consequences if only 80 per cent of people comply with it, and different consequences again if the compliance rate is only 40 per cent. And our obligations under any rule apparently vary, depending on how many other people comply with it. The fact that it would be nice if everyone in my society voted, or paid their taxes in full, does not show that I am morally obliged to vote or pay my taxes even if everyone else is a non-voting tax-dodger.

There are structural similarities between this problem for rule utilitarianism and the "moral weightlifting" problem that I noted for virtue ethics in §8.5. In both cases we have an idealized account of right action that fails to apply to actual agents because a feature of reality has been idealized away: the fact that actual agents are never *fully* virtuous in the case

of the "moral weightlifting" objection; the fact that 100 per cent compliance with any code of rules is almost unheard of in the case of the present objection. One possible response (the "variable-rate" response; see Ridge 2007) is to treat rule utilitarianism as a *disjunction* of codes. We can say, in effect: here is the code that it prescribes given 80 per cent compliance, here is the code it prescribes given 70 per cent compliance, and so on; together with a general injunction to promote conformity to the 100 per cent code, or at least the highest-compliance code possible.

There is an obvious analogy between this variable-rate rule utilitarianism and the version of the virtue ethical account of rightness at which we eventually arrived in Chapter 8, which specifies what to do if you are a minimally virtuous or approximately virtuous rather than a fully virtuous agent, but also tells you to aim to become more virtuous. However – and here the analogy with virtue ethics runs out – this "variable-rate" proposal also raises fairly daunting epistemic problems. To formulate the variable-rate version of rule utilitarianism, we do not just need to know the consequences of *one* level of compliance with *one* set of rules. That might be hard enough; but what we need to know is the consequences of *every* level of compliance with *as many different sets of rules as turn out to be needed to cover all levels of compliance*, from 0 per cent to 100 per cent. To go beyond formulating variable-rate rule utilitarianism and actually implement it, we need to know still more: not only what we need to know to formulate the position, but also what level of compliance we are actually dealing with in the real world now. This sets up the worry that variable-rate rule utilitarianism contains too many unknowns for us to have much idea what verdicts on real decisions that theory might offer.

I turn to a different question that we might raise about rule utilitarianism: why *rules* particularly? There is a possible indirect utilitarian position that we might call virtues utilitarianism, which says this:

> *Virtue utilitarianism:* An action is right iff it is an action in accordance with the set of virtues that, if they were generally accepted in society, would produce the best consequences.

And there is another called motive utilitarianism, which says this:

> *Motive utilitarianism:* An action is right iff it is an action in accordance with the set of emotions that, if they were generally felt by

people in our society, would produce the best consequences (see Adams 1976).

The question "Why rules particularly?" can be read as a challenge to justify the indirect utilitarian's usual focus on optimific rules rather than, say, virtues or motives or emotions (or other entities such as responses or choices or deliberations or laws or institutions; cf. §9.1). It can also be read as a plea, not for any sort of indirect utilitarianism, but for a *global* utilitarianism: maybe the actions to look for are the ones that are based on *everything* that is optimific.

Contemporary debate about this problem connects with a different objection to rule utilitarianism, which usually comes from other utilitarians. This objection – the *collapse* objection to rule utilitarianism – focuses on its *indirectness*. The whole point of utilitarianism, this sort of critic will say, is to maximize the goodness of outcomes. So does rule utilitarianism's focus on general rules (such as "Do not steal"), rather than on particular decisions (such as whether or not to steal here and now), have that effect? Does accepting rule utilitarianism's account of rightness maximize the goodness of outcomes?

The question poses a dilemma for rule utilitarianism. Suppose the rule utilitarian's focus on general rules such as "Do not steal" *does* maximize the goodness of outcomes. But in that case, the rules that the rule utilitarian tells us to keep are rules that the act utilitarian tells us to keep too. For act utilitarianism tells us to do whatever maximizes the goodness of outcomes. If that includes acting as we would if we were obeying rules such as "Do not steal", then act utilitarianism tells us to act that way. And then rule utilitarianism turns out to be just a complicated and roundabout way of saying what act utilitarianism says more simply. As people say, rule utilitarianism *collapses* into act utilitarianism (see Lyons 1965).

Suppose, alternatively, that the rule utilitarian's focus on general rules such as "Do not steal" *does not* maximize the goodness of outcomes, because it leads us to lose utility that we could get by breaking the rule against stealing every now and then. If, by contrast, we accepted act utilitarianism, then we would break the rule against stealing whenever there was more utility to be gained by breaking that rule than by keeping it. So the only distinctive effect of rule utilitarianism on our choices will be to lose us utility. The only differences between the verdicts given

by rule and by act utilitarianism, in practice, will come in those cases where rule utilitarianism tells us not to steal, even though there is more utility in stealing than in not stealing. Therefore rule utilitarianism is bound to be a theory the accepting of which produces less utility than the accepting of act utilitarianism does. So given the choice between the two theories, we have a good utilitarian reason to prefer act utilitarianism: namely, that accepting it produces better consequences than accepting rule utilitarianism.

The rule utilitarian has two possible responses to this objection. One is to abandon maximalism. The rule utilitarian who takes this line will accept that rule utilitarianism is sub-optimal: it does not produce as much utility as act utilitarianism. However, he will say, that loss of utility is worthwhile, because it means that rule utilitarianism can reflect our intuitions about situations where we ought not to do what promotes utility, for example undetectable theft or murder, better than act utilitarianism can. (This is one of the things that Hooker [2003] says about the objection; in this sense Hooker is not what I mean by a consequentialist, either.)

Alternatively, the rule utilitarian can say that his theory is not sub-optimal at all. He can argue that any serious attempt to advocate and practise an act utilitarian moral theory will lead to all sorts of unintended negative effects – loss of trust due to optimific stealing, for example – and that the best way to avoid these unintended negatives is to advocate or practise not act utilitarianism, but rule utilitarianism. In other words, rule utilitarianism is the best way, in practice, of aiming at the act utilitarian target. Or as we might also put it, the best reason for being a rule utilitarian is that act utilitarianism is true, but *unthinkably* true.

This point about unthinkability brings us back to the notion of modularity with which I began this chapter. In "two-level" utilitarian theories (as they are called), there is, typically, a set of rules, values or aspirations close to those found in common-sense morality, which is defended by appeal to a deeper theory that justifies these rules or values on utilitarian grounds. The various species of indirect utilitarianism are the results of this sort of thinking, rule utilitarianism being the best known of these theories. The whole point of such "two-level" theories is to introduce into utilitarianism something like the modularity that the virtues introduce into virtue ethics (see §8.3). It is easy to justify such

modularity on utilitarian grounds in *general* terms. After all, as I argued in §9.5, there is a problem about how there can be any finite deliberation without such modularity. However, two problems then follow.

First, the utilitarian theory that does the justifying sinks into the background. The common-sense rules that have been justified become the main feature of the theory, and it becomes unclear how exactly the resulting theory counts as a *utilitarian* theory, especially when, as is often the case, that theory contains views or attitudes, for example about rule-keeping or the nature of friendship, with an explicitly anti-utilitarian content.

Secondly, the utilitarian faces a real difficulty in attempting to justify any specific form of the modularity that he wants. As we saw in §9.4, any distinctively *utilitarian* justification of this modularity would have, itself, to be a non-modular justification. That is, it would have to be a justification that showed that the preferred form of modularity was to be preferred *because it was literally the best version of modularity available*. But the whole reason we are looking for modularity in the first place is that (as I argued against the ethical rationalist in §8.3) non-modular justifications of this form are very difficult, if not impossible, to sustain.

This means that utilitarianism faces a difficulty about how to combine two of its central ambitions. One is to keep actual deliberative practice manageably finite and definite; the other is to give actual deliberative practice the sort of unique rational authority that could only come from knowing, of any verdict of actual deliberative practice, that it was a literally maximizing verdict.

I shall say more about this difficulty for utilitarianism (and other theories that face the same problem) in Chapter 11. In the meantime, these difficulties about utilitarianism might lead us to attempt to develop an ethical theory on a quite different basis: not on the basis of the promotion of well-being that utilitarianism starts from, but from an examination of the content of the notion of practical reason. This latter line is the one followed by the Kantian. I turn to it in Chapter 10.