The Sanctions of Utilitarianism[1]

ROSS HARRISON

LET ME START WITH A QUOTATION. See if you can place this. 'The question is often asked, and properly so, in regard to any supposed moral standard—What is its sanction? what are the motives to obey it? or more specifically, what is the source of its obligation? whence does it derive its binding force?' Now, given the context of this collection, if you didn't recognise the quotation, you might naturally have supposed that it was Sidgwick speaking. If so, you would have been wrong. It is in fact the first words of Chapter 3 of J. S. Mill's work, *Utilitarianism*, and the title of that chapter is 'On the ultimate sanction of the principle of utility'. So this is where Mill is concerned with what he calls the sanctions of utilitarianism; that is, with the topic of this paper. Now that you have located (or been confirmed in your knowledge) that it was Mill, you may well be thinking that, of course, it could not possibly have been Sidgwick. For Sidgwick was an internalist about moral motivation; that is, once we have identified the right thing to do, there need be no further question about how we are motivated to do it. As Sidgwick puts it in the 'Ethical Judgments' chapter of *The Methods of Ethics*, 'when I speak of the cognition or judgment that "X ought to be done" ... as a dictate or precept of reason ... I imply that in rational beings as such this cognition gives a motive or impulse to action' (VII, 34). So it would seem that Sidgwick could not have a problem about

[1] References to Sidgwick's *Methods of Ethics* give the number of the edition referred to in roman capital letters followed by the page number in that edition.

Proceedings of the British Academy, **109**, 93–116.

moral motive, and hence would not have worried about the sanctions of morality, or utilitarianism. Unlike Mill, it would seem that he could never have written a chapter about it.

This, however, would be too hasty. For the last chapter of the first edition of Sidgwick's *The Methods of Ethics* is entitled 'The sanctions of utilitarianism'; and it is from the title of Sidgwick's famous last chapter that I have lifted the title of this paper. Therefore Sidgwick as well as Mill discusses the sanctions of utilitarianism, and I want to examine this discussion both to investigate Sidgwick's relations to his utilitarian predecessors and also because it may cast light from an unusual direction on the famous end of the first edition of *The Methods of Ethics*. This is where Sidgwick gets caught in the dualism of practical reason and hence finds his whole work a self-confessed failure. If this dualism cannot be solved, then, as he puts it here, 'the Cosmos of Duty is thus really reduced to a Chaos: and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been fore-doomed to inevitable failure' (I, 473).

Three years later, in his second edition, Sidgwick changed the chapter, dropping the title, the beginning, and the end. However, let us continue for the moment to look at it in its first-edition form, where it is a chapter which starts with the title 'The sanctions of utilitarianism' and ends with the word 'failure'. Sidgwick is, typically, more hesitant than Mill. Mill starts his chapter, as we saw, by saying that the question about sanctions is a proper question. Sidgwick starts his chapter by saying that 'We have now, perhaps, obtained a sufficiently clear outline of the manner in which a consistent Utilitarian will behave. But many persons will still feel that, after all, it has not really been shown why a man should be a consistent Utilitarian.' He then remarks that in an earlier chapter

> we seem to have proved ... that it is reasonable to take the Greatest Happiness of the Greatest Number as the ultimate end of action. But in order that this proof may have any practical effect, a man must have a certain impulse to do what is reasonable as such: and many persons will say—and probably with truth—that if such a wish exists in them at all it is feeble in comparison with other impulses: and that they require some much stronger inducement to do what is right than this highly abstract and refined desire.

That he has proved the truth of utilitarianism might well seem to be a sufficient answer to the question of why people should be utilitarians. Nevertheless Sidgwick gets involved with 'inducements', that is, with sanctions. Reasons have to be effective. Yet his interpolated 'and probably with truth' lacks the strength of Mill's brisk 'and properly so'. It remains somewhat open how seriously Sidgwick himself is involved with the problem of sanctions (as opposed to something which 'many persons will say').

Sanctions are referred to in fact much earlier in Sidgwick's great work, and also in all of its editions. If we go back to Chapter 5 of Book 2, we find a chapter which is called, in all editions, 'Happiness and duty'. Here Sidgwick talks about sanctions and in doing so he refers in all editions to Bentham. In the last editions he says, 'here it will be convenient to adopt with some modification the terminology of Bentham; and to regard the pleasures consequent on the conformity to moral rules, and the pains consequent on their violation, as the "sanctions" of these rules' (VII, 164). On terminology Sidgwick is correct. The language of 'sanctions' is pure Bentham. Chapter III of Bentham's major work, his *Introduction to the Principles of Morals and Legislation*, which is the first chapter after his statement and defence of the principle of utility, is entitled 'Of the four sanctions or sources of pain and pleasure'. So now we have another sanctions chapter; that is, all three members of the Holy Trinity of English Utilitarianism, Bentham, Mill, and Sidgwick, wrote a chapter on its sanctions. The four sanctions Bentham lists here are here called the 'physical'; the 'political'; the 'moral or popular'; and the 'religious' sanctions. Later, for example in his map of all possible motives called *The Springs of Action Tables*, Bentham added a fifth sanction, the sympathetic sanction. But the strongest and most important sanctions are the four he cites in the *Introduction* and throughout his life. In each case the name indicates the source of the motivating pleasure and pain. The physical sanction is the pure physical consequences of an action, so that my anticipation of a hangover may be a sanction or motive controlling the amount I drink. 'Political' for Bentham means a legal sanction, that is penalties artificially attached by law to kinds of action hence forming additional motives for people not to do them. So, to stay with drunkenness, there

might be laws against drunkenness threatening fines or imprisonment for being drunk, and hence giving a motive for not drinking. By 'moral or popular' Bentham means public opinion, that is the inconveniences into which I would run by doing those things which are publicly disapproved, such as being drunk in the street. Bentham calls this the 'moral' sanction, but, as Mill later sniffily commented on and criticised (in his 'Bentham' essay), Bentham does not take morality itself to provide a sanction. The incentives are not, that is, taken to come from one's own moral sense but, rather, from other people's. Lastly Bentham has the 'religious' sanction, which is the penalties annexed to actions by the divine law-giver, so if God were to tell me, for example, that drunkenness leads to hell-fire, this would be an additional sanction against drinking too much; it would, that is, give me another reason for not getting drunk.

This is Bentham, whom as we have seen is woven into Sidgwick's text. Bentham particularly appears in the first (as I say, in the *Introduction*), more historically oriented, edition and here Bentham is much admired. When, for example, Sidgwick dismisses Bentham's posthumous *Deontology* as really being a work of Bentham's own pen, he does so because he finds in it things which, as he puts it, are 'impossible to attribute to so exact and coherent a thinker as Bentham' (II, 68). So Bentham is taken to be exact and coherent. Sidgwick is also exact, even if, attempting a wider range, he is not quite so coherent. He is particularly exact, I think, in the much longer passage which in the earlier editions stands in the place of the short summary about Bentham and sanctions which I quoted. Here Sidgwick says,

> It has been already observed, that while stating General Happiness as the right and proper end of conduct, Bentham still regarded it as natural and normal for each agent to aim at his own individual happiness. He therefore considered human pleasure (and pain as its negative quantity) from two quite distinct points of view: first as constituting the end and standard of right conduct, and so determining the rules which Bentham and other rational philanthropists would desire to be generally obeyed in any community: and secondly as constituting the motives (whether pleasures or pains) by which each member of the community is or may be induced to conform to these rules. (I, 148)

Then he starts to classify what he calls 'these Motives or Sanctions' in a way which is also common to the later editions.

I think that this is exactly right as a description of Bentham. The fact that, as Sidgwick elsewhere puts it, 'there is ... in Bentham's mind no confusion and no logical connexion between his psychological generalization and his ethical assumption' is both crucial and also frequently overlooked. We can see the distinction at the start of Bentham's 'Sanctions' chapter where Bentham says 'Having taken a general view of these two grand objects (viz. pleasure, and what comes to the same thing, immunity from pain) in the character of *final* causes; it will be necessary to take a view of pleasure and pain itself, in the character of *efficient* causes, or means' (III, 1). In other words, after showing what ought to be done in his chapters on utility (the final causes) Bentham now turns in his 'Sanctions' chapter to how it may be done. As well as ends, there are means; as well as what Sidgwick called in that long quoted passage the 'end and standard of right conduct' there is what Sidgwick called there the 'motives ... by which each member of the community is or may be induced to conform to these rules'.

Indeed, Bentham's whole project only makes sense if such a fundamental distinction between psychology and ethics is made. For Bentham's project is precisely to take people as they actually are and then to see what system of government and legislation is required to make them do what they ought to do. The idea is, as Bentham puts it, to 'promote the happiness of the society, by punishing and rewarding' (*Introduction*, VII, 1). That is, the correct evaluative end, which is the happiness of society, is to be provided by appealing to people's self-interested motives, using threats of pain or hopes of pleasure, that is, using punishment and reward. The legislator therefore needs to know both the value theory and also the psychological reality of people in order to know which sanctions should be applied to which people to get them to do which things. General happiness comes from people not stealing, so punishments are fixed for theft, forming a sanction against bad behaviour. The same applies in Bentham to the proper structure of organisations and, indeed, the construction of government itself. Bentham's panopticon prison, to take a famous example, is precisely meant to be a physical or spatial solution to the question of how the

self-interested prisoners are to be motivated to do what they ought; but the same applies to the problem of who guards its guards, the principles of management by which its self-interested governor is to be motivated to run the prison properly.

What we get in Bentham, then, is a political solution to a moral problem. Theft is bad; people do not do as they ought. The solution is to have law and government, which by imposition of the so-called 'political' sanction makes it in people's interests not to steal. Other sanctions are involved, for as Bentham says, the political sanction involves the physical; the physical walls of the prison are part of the deterrent for bad behaviour. Also, as Bentham says in his 'Sanctions' chapter, the legislator overlooks the religious and social sanction at his peril. And as well as what Bentham calls direct legislation, there is what he calls indirect legislation, that is, the other ways than punishment by which a legislator can influence or educate people. However, the whole work is written from the perspective of the legislator.

It is therefore a political solution to a moral problem. It is what a legislator does; or, more accurately, what a good legislator should do. However, the present question, or the question of Mill's and Sidgwick's 'Sanctions' chapters, is whether the same can be done for morality; done for morality, that is, without using legislation. Where the texts of the early and late editions join, Sidgwick next says that the 'sanctions we may classify as External and Internal'. External is like Bentham; and Sidgwick here identifies what he calls 'Legal Sanctions' and also 'Social Sanctions'. (So Sidgwick's 'legal' corresponds to Bentham's 'legal or political', and Sidgwick's 'social' corresponds to Bentham's 'popular or moral'.) However, when Sidgwick mentions 'internal' sanctions, he becomes more like Mill than Bentham. In Mill's 'Sanctions' chapter, with which I started, he also says that 'sanctions are either external or internal' and indeed he spends much more time on the internal sanctions than the external, that is on how I feel when I do wrong and so on. Now this might be thought to be the clue we need as to how we can find sanctions for morality. As well as the pains of the externally imposed law we have the pains of the inner conscience. Sidgwick, for example, notes here that 'The internal sanctions of duty ... will lie in the pleasurable emotion attending virtuous action, or in

the absence of remorse' (VII, 164). This might seem like the clue. But in fact internal sanctions do not make the problem different in any fundamental way. Bentham may have been criticised by Mill for leaving them out, but as long as the pains of conscience are taken as merely unpleasant feelings, like having a bad stomach ache, they operate in the same way as the externals: the anticipated hangover may stop me drinking; the anticipated misery of remorse may stop me breaking my promise.

Another way of putting the point is to say that there is in Bentham no problem about moral agency. All agency for him is purely self-interested and so the solution to the problem of morality is purely political. Once the political is taken away, it is not clear what a purely moral problem and solution would be. We would just have people acting, and by luck or divine intervention happening or otherwise unintentionally to hit the right target. However, at least unless we are God, there is nothing that can be done about it. And even if we are God (or introduce God), we have another politician, a divine legislator, who, by keeping the home fires of hell burning bright, manages to get people to do the right thing.

At first sight, therefore, there could not be for Bentham a purely moral analogue to the problems to be solved by legislation. Legislation is an art, a form of cooking in which the right cake is to be made with these ingredients. But, take away the political machinery and there seems nothing left to do; no comparable purely moral task. However, even on Bentham's own account, this cannot be quite right. As well as what the cooks cook, there is the question why they cook what they do. In other words we have the question of the legislators' own motives for action. Why are they attempting to achieve the greatest happiness of the greatest number; and if they are, are they not being moral in a way that the psychology cannot explain? The account is of what a legislator should do, and this does not make sense unless there is scope for the legislator to make morally motivated choices. There is also the problem of Bentham's own position, that is of the philanthropic adviser to the legislator who explains how the legislator should act to get general happiness. Bentham himself seems to be acting in a morally motivated manner, for, as Mill pointed out in his 'Bentham' essay, Bentham was

himself very unlike the self-interested operators whom he took as the typical or universal specimens of human activity.

There do therefore seem to be exceptions to Bentham's psychological claim that people are animals which act in a universally self-interested manner. Bentham in fact sometimes allows this. And Sidgwick in the long comment which I quoted says merely that Bentham regarded it as 'natural and normal' for each agent to aim at their own happiness. However, Bentham did usually claim it as a universal truth (and is quoted by Sidgwick elsewhere as doing so). And when Sidgwick produced the second edition of *The Methods of Ethics*, he left the long passage I quoted alone except that for 'natural and normal' he substituted 'every human agent actually does aim at his own individual happiness' (II, 148). However, either way, it does not spoil Bentham's legislative project, his cook's task, because for this, knowledge of general rather than of universal behaviour is sufficient. As long as you know that in general people do not want to go to prison, you can institute this as a sanction to deter theft and its efficacy won't be much undermined by occasional monkish characters who find prison an answer to their spiritual needs.

Bentham still has to give some explanation of why he himself (apparently altruistically) advises legislators and also why any legislator (apparently altruistically) should listen to the advice. As for himself, his explanation is that he just happens to be one of those people who are not motivated in a narrowly self-interested sense; or (the trivial form of this) who get their happiness from doing good to others. As for the legislator he is attempting to advise, presumably the strategy is just to wait for one which happens, by the same exceptional chance, to look benevolent. Then when you get her, a Catherine the Great of Russia perhaps, you head for Russia and start writing advice.

This may not be inconsistent. However, it is very risky. Benevolent dictators are still dictators, and if they institute the perfect system of law on Monday they may still hang you without trial on Friday. So a better answer is the one which Bentham came up with later in his life. This is to have a political system which will automatically, in its normal process of running, produce the good. The system is democracy. Here, again, we put together the separate psychological and ethical

principles to achieve the result. The ethical principles state that the consequence which ought to emerge is the greatest happiness of the greatest number. The psychological principles state that people in fact aim at their own happiness. Therefore the greatest number will in fact aim at the happiness of the greatest number. Therefore putting the greatest number in charge (as happens in majoritarian democracy) means that people following their normal psychological courses will happen to come up with the goods. The moral problem of getting the right thing done is again solved politically; the political system ensures success.

We have now got rid of the legislator who has to be heroically moral, or otherwise has uncertain motivations. Merely moral motivation is no longer required. Instead the sanctions are provided by a machine which runs by itself, just as Bentham's panopticon prison is designed as a machine to run by itself, a machine to produce moral good without moral motivation. Normal psychological sanctions are sufficient and the job is all done by possible pains. This is again a political solution. However, it now seems that we could describe something similar for morality without involving the political. For when we have a machine that runs by itself we no longer have the cooks. The Benthamite philosopher is no longer an adviser on how to bake cakes but merely a commentator on the fact that the machine in its normal workings seems to be producing the good.

However, if mere commentary is all that is possible, then it would seem that exactly the same could be done for morals independently of politics. It could be pointed out in a precisely analogous way that the workings of the normal human psychology happen by and large to produce the right moral consequences. This could be taken to have divine backing, as with the theological utilitarians from Cumberland to Paley. The benevolent God is taken to have so fixed our psychology that (by and large) acting on our normal impulses means that we end up with the greatest good. This can also be given an invisible-handed economic spin whereby God (no doubt with time off after fixing physics) benevolently fixed the truth of general equilibrium theory so that satisfying our individual desires is again all for the best. However, even with the ruthlessly secular utilitarians such as Mill (and Spencer or Leslie

Stephen) the long-run experience of evolving humanity has given us a useful nature, so that by following our natural desires we are happening also to be producing general good. So we reach again Mill's 'Sanctions' chapter. But so also, it would seem, we may reach Sidgwick's 'Sanctions' chapter. We may, that is, now have this chapter in precisely the way it is, starting with sanctions and ending with the dualism of practical reason.

For the question now is whether people acting self-interestedly nevertheless happen to produce the right moral consequences. And, as long as there is a large overlap between what you will arrive at if you act self-interestedly and what you will arrive at if you act benevolently, then there are indeed sanctions to be moral. And such an overlap is precisely something that Sidgwick argues for in this chapter. Even if people are normally or universally the self-interested machines portrayed by Bentham, nevertheless they will still do those things which as a consequence produce the right results. They will, that is, produce the same things as they would have produced if they had instead been acting on benevolent, universal, reasons. It is exactly the same as the democratic story: morality is the unintended consequence of the normal self-interested working of the machine. Of course it is not automatic or guaranteed. Occasionally there will be divergence. But on the whole the machine works. The cakes are made but nobody intends to cook.

Alternatively, we might be less sanguine about there being a large overlap between what naturally arises from self-interested motivation and what would arise from universal benevolence. However, it would seem that we could still put the point as follows. Either self-interested and benevolent reasons agree in their practical effects or they do not. If they agree, then we have self-interested reasons to be moral, hence solving the sanctions of utilitarianism problem, and hence also showing that the supposed problem of dualism of reason is merely apparent. Alternatively, they do not agree. Then we are into problems on both counts. Sanctions cannot now be reliably presented for being utilitarian and we are also into deep difficulties over the dualism. Sidgwick points out in the course of the chapter that although there is normal coincidence, there is both possible and actual divergence. For him, doing the

right thing may not be in our interest or make us happy, whereas, as he puts it in a footnote, 'some few thoroughly selfish persons appear at least to be happier than most of the unselfish' (1, 464n). So morality may well not be good for you. The course is set for the final failure whereby the cosmos of duty cracks apart into chaos; but this failure will also be a failure for the topic with which the chapter starts, that is with finding the sanctions of utilitarianism.

All this would seem to support Sidgwick's apparent strategy of writing a chapter about sanctions and ending up discussing the dualism of the practical reason. However, it is not in fact this simple; and it is not simple for reasons which Sidgwick himself brought out in his discussion of what he calls psychological hedonism, that is, the psychology of both Bentham and Mill. For Sidgwick there are two things wrong with Mill's famous proof of utilitarianism: the premiss and the inference. The premiss is that happiness is the only thing aimed at as an ultimate end; the inference is that therefore this is the right and proper end of conduct, the thing at which we ought to be aiming. The criticism of the inference, of the move from *is* to *ought*, is a criticism of a mistake which Sidgwick thinks Mill made but which, as we have seen, he thinks that Bentham never made, holding as Bentham does the descriptive psychology quite distinct from the evaluative ethics. However, both Bentham and Mill share what I have called the premiss of Mill's argument, and this also Sidgwick criticises. Sidgwick points out that people do not only aim at their own pleasure. People may, for example, be concerned with things which happen after their death, when they are no longer around to experience anything or have any pleasure. More generally, Sidgwick shows that desire is not necessarily a desire for pleasure. My hunger makes me desire food, but this desire is not a desire for the pleasure which the food may give. Indeed, in what Sidgwick calls the paradox of hedonism, it may be that the best way to get pleasure is not explicitly to pursue it. What all this means is that the things which Bentham holds together come apart: that is, desire, motivation, anticipation of pleasure, sanctions.

The title of Bentham's 'Sanctions' chapter was the 'sanctions or sources of pain and pleasure'. We have just seen that sanctions, that is Bentham's pleasure and pain, are not the only motives to action. But even

if they were, or when they are, how on the Benthamite account are they meant to work? That is, how is the pain supposed to connect with the action? A present pain may explain present action; but most of the action supposedly explained by sanctions is explained by possible future pains, such as threats of punishment. Here it is not the fact of the future pain that is meant to explain motivation and action but, rather, my anticipation of it. Unknown pains will not motivate, yet I can be motivated by false expectations. Hence it is not the pleasures and pains in themselves that do the work but how they seem to me. But these may diverge; I can make mistakes. One example of this, which Bentham notices, but which is particularly important for Sidgwick, is in my estimation of the value of future pains. Our psychological practice is to discount them, so that the further off they are in the future, the less importance we give to them. Hence, in his 'Value' chapter of the *Introduction*, Bentham has 'propinquity or remoteness' as one of the measures of value (IV, 2), and this is correct with respect to the descriptive psychology which he is laying out at this point as to how things are actually valued; that is, the effect they actually have on us. However, Sidgwick does not think that this is rational, even from a self-interested, egoistic, point of view. What we should be doing is displaying no time-preference at all since the value of a state of affairs should be the same at whatever time it happens. (We can discount for certainty, and, normally, the further off in the future something is, the less certain it is; however, this is a different reason for discount and one mentioned separately by Bentham.) So now we get another crack in the system, in the things which were held together in the earlier utilitarian psychology. This time it is between self-interest on the one hand and motivation by apparent pain and pleasure on the other. Even if we think that there is nothing to self-interest other than one's own happiness, and even if we think that happiness consists of no more than pleasure and the absence of pain, it will still be the case that the motivating force of apparent pleasures and pains may not be in our interest. Even if we anticipate future pleasures and pains correctly, we discount them more than we should; hence we act too much on the appearances and not in our long-term interest.

The importance of this in the present context is the following. If we are to be rational egoists, then we should not follow the immediately

motivating force of apparent pain and pleasure. But sanctions are what actually motivate; that is, what actually explain the psychology of action. Hence we cannot identify sanctions and rational self-interest. Hence, contrary to first appearances, an account of the sanctions of utilitarianism will not be the same as an account of the coincidence of self-interest and benevolence. It may be remembered that when Sidgwick introduces the question at the start of the chapter, he talks of how 'a man must have a certain impulse to do what is reasonable'. Just so. However, now we find that to be 'reasonable' includes both rational self-interest and also rational benevolence. Someone may indeed need an impulse in addition to the mere perception of what is reasonable. But, if so, this applies equally to both ways of being reasonable. That is, we need an additional impulse to be rational in terms of our self-interest as much as we do in terms of benevolence. The question of sanctions, therefore, if it applies at all, applies equally to both. We have to explain how people are able to be self-interested as much as we have to explain how they are able to be benevolent. The question of sanctions for being rational cannot therefore be the same question as whether there are self-interested reasons to be moral.

Sidgwick is quite explicit that in this he is following Butler, another past philosopher with whom he is in conversation. It is Butler who, as Sidgwick quite explicitly says, gives him the dualism. It is Butler who, before Sidgwick, brings out how people fail to act in their own interests. Butler is a clergyman giving sermons. We might therefore expect him to criticise self-interested actions and tell people instead to be moral. However, perhaps surprisingly, he criticises them instead for failures of self-interest. As he puts it in his *First Sermon*, 'men in fact and as often contradict that *part* of their nature which respects *self*, and which leads them to their *own private* good and happiness; as they contradict that *part* of it which respects *society, and* tends to the *public* good'. In other words, we are equally deficient in both of Sidgwick's ultimate reasons (rational self-interest; rational benevolence), and both equally need strengthening.

Such strengthening might be by political action, again forming a political solution to a moral (or here it would be better to say a rational) problem. Hence in contemporary road-traffic legislation, the

government may act to control speed and safety of vehicles, aiming at the benevolent effect on the other people who would otherwise be damaged by the driver's actions. But it can equally intervene politically, not allowing people to drive without their seat belts, or motorcyclists to ride without crash helmets, thus instituting sanctions designed to make people more concerned with their own future happiness, making them more rational egoists. Similarly for the areas of indirect legislation or advertising, whether public or private. The pictures of the far-off famine make the pain of others vivid and real and so motivate people into benevolent action. Similarly, the far-off sufferings of lung cancer may be made vivid and real to present-day smokers, motivating them to act in their own interests. Again it is not the fact of the pleasure and pain which motivates, or gives the sanction, but, rather, how it is made apparent to someone at a particular time and place.

We can, of course, as philosophical or psychological commentators, attempt to give an external description of how it works. An explanation, that is, of how it is possible for people to be rationally benevolent or rationally self-interested; of how the show is kept on the road. Even if we disagree with Mill's remark that 'desiring a thing and finding it pleasant, aversion to it and thinking of it as painful, are phenomena entirely inseparable' (*Utilitarianism*, IV), we may still recognise that there is a close connection between finding something painful and the desire to avoid it. We recognise that pains are naturally motivating and that the prospect of pain gives reasons for action. So, if we are concerned about how the show is kept on the road, how people may be successfully self-interested or how a society is able to reproduce itself in its moral culture, then the explanation will be helped if we can see how doing the moral things is generally pleasurable and not doing them is generally painful, just as we can explain the physical reproduction of society by the pleasures of conception.

This may take us some way. However, pains and, even more, the more amorphous so-called pleasures are merely some motives among others. So if we are explaining performance by motivation they will only take us, at best, part of the way. We might expect a general fit, but there is no reason why it should be perfect. This, as we have seen, is also true of prudence, so that following our immediate desires or our

anticipations of pleasure or pain may be anything but prudent. Pain is, no doubt, an important signal. When our hand on the stove pains us, it is likely also to be the case that not just pleasure but also self-interest recommends its removal. However, when Emily Brontë returned from a walk during which she had been bitten by a rabid dog and cauterised the wound with the red-hot poker from the fire, she was behaving prudentially and rationally. It was a heroic piece of self-interested self-sacrifice.

Psychology is a normative science. We interpret people not just as believers of the true but also as lovers of the good. In understanding (or interpreting) them, we understand how they would justify themselves both to themselves and also to us; that is, what they would recognise as ultimate reasons for action. As such, Sidgwick's (and before him Butler's) claim seems to me to be highly plausible; namely that we recognise as such ultimate reasons that something is in someone's interest (or leads to their happiness) and also that something is in the general interest (or benefits someone else). An action is explained if it is shown how it avoids pain either for the agent or for someone else. It explains it by providing a recognisable justification. As Sidgwick says, 'happiness appears to be a reasonable end ... if I can say of any action that it makes me happier, it seems that no further account need be given of my doing it' (I, 59).

Hence we have the dualism, but here in a different way from anything directly connected with sanctions. If 'sanctions' just means what can (rationally) motivate me, then either of these considerations can provide sanctions, that is, reasons. It is not that one of them (say, self-interest) forms a particular kind of sanction different from the other. On the other hand, as we have seen, if 'sanction' means an immediate anticipation of pleasure and pain, then neither reason will directly follow from such sanctions. So the asymmetry which seems to be implied in Sidgwick's last chapter between self-interest and benevolence drops away. Indeed, it has to drop away if the dualism of practical reason is to assume the important and final position in which Sidgwick places it, for such dualism necessarily depends upon the equal weight and value of each of its two elements.

An important part of the answer to why we are able to be moral may be no better than that people are able to see the truth (or, alternatively,

that the truth is constituted by what people are able to see). Take the comparable question of how we are able to do mathematics; that is, what keeps the mathematical show on the road. This is both a theoretical and a practical question. It is the question of how we can go on continuing the series correctly, adding plus 2 in the same way as everyone else when we reach 115,326; 115,328; 115,330 ... However, it is also the question how, having 4 plus 1 bolts, we know that we need 3 plus 2 nuts to fasten them. We could explore how the size or workings of the brain enables us to do this. We could look at education and initial conditioning. But in the end we may get a no more interesting answer than that it is true that 115,326 plus 2 is 115,328 or that it is true that $4 + 1 = 2 + 3$. It is a truth which, as it happens, we are able to see; or, alternatively, we see it, and truth here is constituted by what we are able to see.

In the more obviously practical, or moral, case it need be no different. The show is kept on the road; people are able to tell that they should not torture innocent children. Again, no doubt, we could investigate brain power or initial education. But, again, the best answer may be the banal one that we can do it because it is true and we are able to see it; or, again, alternatively, the truth here is constituted by what we are able to see. That is, we are motivated to be moral by seeing that what is moral motivates. Of course if we thought that only self-interest motivates, then we would have to see how this could give a reason for so-called 'moral' actions, for keeping promises, not torturing babies and so on. However, this is not Sidgwick's problem, for Sidgwick thinks that as well as the ultimate rationality of self-interest there is the ultimate rationality of benevolence. Hence I have a direct practical reason to be moral; and having a practical reason is, for Sidgwick, to have something which motivates. So all that is required for him to be able to explain why we are able to be moral is to show, firstly, that this reason can be demonstrated to be true and, secondly, that it is a truth of which people are aware.

Sidgwick does, I take it, prove both of these things in *The Methods of Ethics*. He does it by relating utilitarianism (or rational benevolence) in two different ways to the doctrines which he labels intuitionism. We have what he calls philosophical intuitionism, the high-level intuitive

(or rational) principles epitomised by Kant. From these it follows (or so, at least, Sidgwick thinks) that 'the good of any one individual is of no more importance, from the point of view ... of the Universe than the good of any other' (VII, 382). That is, good should be considered impartially; hence the rationality of benevolence, the rationality of considering another's good as neither more nor less important than your own. However, if we stop here this might be like a truth provable by professors but unavailable to the multitude. Hence the importance of his examination of what he calls common-sense morality; that is the morality of the people. Sidgwick himself is concerned to point out in the second edition preface that this morality is also his own, but the next point follows whether Sidgwick is part of the government house elite or part of the rabble. This point is that for Sidgwick this common-sense morality can be systematised and reduced so that it reveals what he calls 'unconscious utilitarianism'. Hence the rabble, the people (all of us), do know the right (utilitarian) morality in our everyday or common-sense morality about not killing our granny and so on. Sidgwick considers common-sense morality a kind of intuitionism, and the book of the *Methods of Ethics* called 'Intuitionism' is mainly an account and systematisation of common-sense morality. So, for him, both kinds of intuitionism converge on utilitarianism. To show this was, after all, the great feat of the *Methods*. In the present context it means that everyone, whether philosopher with knowledge of the form of the good or slave boy scribbling in the sand, knows the right answer to the question about what they ought to do. And this knowledge is sufficient for both to keep the show on the road. No other sanctions are required.

Of course if the good is dual and possibly diverges, then there may be problems. Earlier Sidgwick himself points out, while he is talking about sanctions, that the sanctions mentioned by Mill and Bentham may diverge. For example, the deliverances of my conscience may not coincide with Bentham's 'popular or moral sanction'; that is, with what public opinion presses on me as right conduct. Divergence is possible and divergence causes practical problems. But any set of things sufficiently simple to be useful in either explanation or justification will cause such occasional practical problems. General lack of divergence is all that is necessary for the set to be practically useful. Sometimes

(because of divergence) there may be a hiccup but, in general, this pair of fundamental reasons for action works perfectly well and works much better than any suggested alternative. This includes the alternative of using just one member of the dualism.

This is, if we treat it as a practical problem, a question of what we should do. However, Sidgwick, because his dualism is of reason rather than of sanctions, thinks that the problem here is more theoretical than practical. For him there cannot be two equally fundamental but possibly competing reasons. He thinks that such possible conflict is fatal to the theoretical enterprise of truly describing practical reason; possible conflict reduces cosmos to chaos. However, I do not myself see that this is a serious, let alone fatal, problem. Early in the *Methods* Sidgwick says that 'no doubt it is a postulate of practical reason, that it must be consistent with itself: and hence we have a strong predisposition to reduce any two methods to unity' (I, 66). However, he also adds 'that it is a special object of the present work to avoid all hasty and premature reconciliations'. At the end, hundreds of pages later, we might think that he had successfully avoided being hasty and that it was time for reconciliation. Yet it is just here that he thinks that he fails in reconciliation.

As a consequentialist ethic, the problem is not that the two reasons give quite different reasons for action, quite different ways of thinking about our intentions when acting. Providing the right thing is done whichever way we think about it, this would not be a problem on a consequentialist ethic. The problem is that, according to how we think about it, different consequences could follow. Otherwise it would just be like two different ways of thinking about the same thing. To take an analogy, if it could be shown that the two ultimate principles necessarily fitted together, then it would be like analytical geometry. If you have a fundamental algebraic turn of mind, then you can understand these truths algebraically, understanding figures like circles and ellipses in terms of their equations. Alternatively, if you have a geometrical turn of mind, you can think of these figures as what the algebra is really about, understanding the equations in terms of the diagrams. There is a dualism of perception, but it can also be shown mathematically that, whichever way we choose to see it, we will come up with

corresponding results. So, applying the analogy, some may see our actions in terms of enlightened self-interest, some in terms of rational benevolence. However, whichever way they are seen, corresponding truths with respect to action will emerge. The same behaviour ensues, even if it is given different descriptions.

From Plato, geometry may be taken to be a description of ideas (of the ideal world). The actual figures we draw in the sand fall short. In ethics, what is described is doubly ideal. It is, again, ideas, ideas which our actual actions in the sands of this world fail to realise. Yet these ideas are here themselves the ideal, the good, the goal; not something which describes the actual world but, rather, something to which we would wish the actual world to conform. We may rationally see that this ideal contains the good for someone of their interests being met. We may also see that the good contains the impersonal meeting of such goods. This practically ideal world, this best of all possible worlds, therefore contains both. We may, I think, consistently suppose that this best of all possible worlds contains both relative and non-relative goods (the good from my perspective and the good in itself). Then, for someone in this best of all possible worlds, it would be like my description of analytical geometry. There, whether you think algebraically or geometrically, you get the same truths. Here, whether you think in terms of relative or non-relative goods, you think that the same things are good. You can pursue your own good, you can pursue the impartial good; what you pursue is the same. This is the best of all possible worlds. The actual world is not like this. However, the actual world would be a better world if doing the right thing (objectively, or non-relatively, speaking) did not involve self-sacrifice. It would be a better world if realising your self in pursuit of your fundamental interests resulted in an objectively better world also for others.

These points can be put in both a relativised and a non-relativised way. If we just think of goods from no point of view, then it will be trivially better (in terms of such non-relativised goods) if we can move from a situation in which someone sacrifices herself to produce good to others to a situation in which she can produce the same good to others without sacrificing herself. For we have the same good for others but more good for her; hence more good overall. However, this is too

simple. We are concerned not with a world which is best in every dimension but, rather, with the best possible world; and for this some sacrifice may be unavoidable. The utilitarian theory of government imposes potential costs on some (by threatening punishment) for the greater good of others. As people are, a perfect system of punishment is the best that is possible.

This is one way in which this account is too simple. It also is too simple in a more fundamental manner. For if we just consider goods from no particular point of view, we miss the centre of Sidgwick's dualism, which necessarily depends upon considering goods in a relativised manner; that is, we have to consider how John's good gives John a special reason to act, over and above it just being a good (otherwise Sidgwick's refutation of egoism would apply). So we need as well a relativised way of making the point that self-sacrifice is not involved in the best of all possible worlds. In fact this can also be done. It is better for me if I can move from a world with a certain amount of good for me to a world in which there is more good for me and the same good for others. Not just better in itself, but better for me in a relativised way. Furthermore, it is also good for me that what is good for me is not incompatible with what is good for all. Conflict in goals is not just bad; it is also bad for me—bad in a relativised way.

Reconciliation between the two fundamental ultimate reasons for action would be premature if it can be shown that they are both ultimate, that both provide fully explicable motivation, or (in this sense) sanctions. Take, for example, the so-called 'golden rule', the precept that you should do unto others as you would that they should do unto you. This was the kind of moral truth which Sidgwick would have learned at his mother's knee, at church, and in school, being enshrined in the Anglican catechism. 'What is thy duty towards my neighbour?', the catechist asked, and the reply to be learned was 'My duty towards my neighbour is to love him as myself and to do to all men as I would they should do unto me.' However, this is not just an Anglican eccentricity. Behind it stand similar statements in both the Old and New Testaments of the Bible. Nor is this just a foible of Western civilisation. As was observed in the anthropologically conscious seventeenth century, tenets like this were discovered to hold in widely dispersed

cultures. Pufendorf, for example, noted it in both Confucius and the Incas. So this worked for them as a posteriori evidence of an a priori, or purely rational, truth. Something seemingly held by all people and without any evidence of acquisition by cultural diffusion must be a truth of reason, perceivable by people just because they were rational animals. Otherwise put, it was inscribed in the hearts of men by the hand of God, so explaining how people without biblical revelation might know the moral truth. It was therefore part of natural law. Indeed, even that highly eccentric proponent of the new natural law, Thomas Hobbes, said that all the laws of nature had been, as he put it, 'contracted into one easy sum, intelligible, even to the meanest capacity; and that is, *Do not that to another, which thou wouldest not have done to thy selfe*' (*Leviathan*, ch. 15). That Hobbes, as Sidgwick points out in *The Methods of Ethics*, is here propounding the rules to be followed for self-preservation just makes the universality of this dictum even more striking.

So here we have a good example of an obvious moral or rational precept; something which gives good reasons for action; something which in the seventeenth century and earlier was inscribed by the hand of God but which later became enshrined in one of its aspects in the high gospel of pure practical reason. Sidgwick himself says (in his 'Sanctions' chapter) that 'I find that I undoubtedly seem to perceive, as clearly and certainly as I see any axiom in Arithmetic or Geometry, that it is "right" and "reasonable", and "the dictate of reason" and "my duty" to treat every man as I think that I myself ought to be treated in precisely similar circumstances' (I, 470). However, what does this intuitively compelling moral truth tell us to do? It contains two points, one about the nature of the good and one about its distribution. On the nature of the good, it gives as a practical rule for identification that the good is things which you would want; or, alternatively, that the bad is things that you would not want. Things that you would not have people do to you are bad things. So this identifies the good (or, alternatively, the bad). Then the other point is about distribution. It is that you should impartially promote this good in others as well as yourself; or not impose this bad on others any more than you would on yourself.

The impartial bit is obvious, recognisably moral, and receives expression in the requirement that moral maxims be universalised. It goes with Sidgwick's 'universal reason', his rational, impartial, benevolence. However, what is equally interesting here is the other part, that is the part which identifies the good rather than saying how it should be distributed. This is done by what people find good for themselves; that is, it is done by the relativised sense of the good in which my good provides me with particular reasons for me to act. Unless the good is identified in this way, then there is no point talking about how you would that people should act unto you. In other words, the formula only makes sense if it is already supposed that we naturally have self-interested reasons for action; that we naturally understand that our good gives a reason to us. Only if we understand this can we use it as a means of identifying the good, and only if we can identify the good can the impartial distribution point operate. Hence, to have the familiar moral rationality of what Sidgwick calls universal reason, we also need the rationality of what Sidgwick calls individual reason. As he says, they are both rational principles.

Sidgwick's problem is not that both of these ultimate principles are deep, but that they do not necessarily converge. They are neither necessarily connected in themselves nor, at least for Sidgwick, is it a necessary truth that there is a god who, as a necessary attribute of his goodness, connects them. So they possibly diverge. However, because these are truths of practical rather than speculative reason, this should not matter. The two may diverge. All this shows is that, if they do, we are not in the best of all possible worlds. The two may diverge, but it can still be true that they ought not to diverge. In the best possible world they would coincide. A god, or indeed any legislator who was both benevolent and omnipotent, would bring this about.

That is how I think that Sidgwick should have got out of his problem. I think that he could have taken the problem on its own terms and still solved it. However, this is not what he actually did. Instead he noted the strong objections to his startling confession of failure and rewrote the chapter. He dropped the famous ending, and so dropped the explicit declaration of failure. But he still held, as he puts it in the second edition preface, that the main discussion of this chapter was

'indispensable to the completeness of the work'. He still quite explicitly affirms both principles and indeed for the first time he describes it as a 'dualism of the practical reason' (II, xii). So the main offensive claim is maintained, and maintaining it means, at least in Sidgwick's eyes, that the hoped-for cosmos of duty is still a chaos and the main project is still a failure.

In fact, these changes make this point more conspicuous. For, as Sidgwick also claims in this new second edition preface, the main misunderstanding he wished to avoid was that his intention was to argue for only one of his three methods (utilitarianism) at the cost of the other two (egoism and intuitionism). With intuitionism, it is true, he effects a reconciliation. However, Sidgwick wishes to make it quite clear that with egoism he has effected neither a reconciliation nor a defeat. Egoism is still fully in play along with utilitarianism, and hence the dualism of reason, or failure to unify.

Other changes which Sidgwick did not remark on in his new preface make this more conspicuous. The last chapter in the first edition is Chapter 6 of Book IV, the book entitled 'Utilitarianism'. So, as such, it looks like the end of the account of utilitarianism, the end of the account of what looks like Sidgwick's preferred doctrine. This is helped by its title, the title I adopted for this paper, 'The sanctions of utilitarianism'. As such it seems to fit neatly after the preceding chapter, 'The method of utilitarianism', which comes after 'The proof of utilitarianism'. The whole forms an exposition and defence of utilitarianism similar to Mill's in content although superior in argumentative texture.

For the second edition Sidgwick dropped this. The last chapter was moved out of Book IV and given a heading of its own, of equal weight to the headings of the previous four books. It now forms a separate end to the whole work rather than being the end of the utilitarian Book IV. He might have called it 'Book V'; instead he calls it 'Concluding Chapter'. He also gives it a new title. Instead of the title referring to utilitarianism, used by this paper, it is henceforth called 'The mutual relations of the Three Methods'. Hence, as he says in the new preface, he makes quite specific that his interest in this chapter is in the relation between utilitarianism and egoism, not in further defence or discussion of utilitarianism itself. This all necessarily makes any tension involved

in the dualism more fully embedded in Sidgwick's overall project, even if perhaps not quite so superficially apparent in the closing words. He holds, as I remarked, that the chapter, and hence the statement of the dualism, was 'indispensable'.

What else might he have done? If he had instead carried on with the utilitarian tack implied by its original heading, he could have dropped the talk of 'sanctions'; indeed, simply dropped the chapter. Then the work would truly have been the defence of utilitarianism which it was naturally read as being and which, in spite of Professor Sidgwick's protests, it still obviously is. The work would not then have ended with the word 'failure'. Instead it would have ended with the concluding words of the previous chapter, which Sidgwick left unchanged from first edition to last. Here he notes the 'stress which Utilitarians are apt to lay on social and political activity of all kinds, and the tendency which Utilitarian ethics has always shown to pass into politics'. He notes, that is, the stress on what I earlier called the political solution of moral problems. And then he concludes the chapter, and it might have been the book, by saying, 'A sincere Utilitarian, therefore, is likely to be an eager politician: but on what principles his political action ought to be determined, it scarcely lies within the scope of this treatise to investigate.' A fine conclusion, and one which makes way for Professor Sidgwick's next treatise, *The Elements of Politics*.