Taking Stock of Utilitarianism

Bernard Williams ended his well-known ‘Critique of Utilitarianism’ with a prediction about the theory: ‘The day cannot be too far off in which we hear no more of it’.[1] That prediction looks more likely to be correct now only because, during the last forty years, humanity has made it more probable that the day is not too far off in which we hear no more of anything. Within philosophical ethics, utilitarianism continues to flourish.

I was initially asked to speak at this conference about what utilitarianism is, and that is what I shall begin by doing, focusing in particular on the version of the view which I believe is common to the so-called ‘classical’ utilitarians: Jeremy Bentham, J.S. Mill, and Henry Sidgwick. I shall then go on to consider how plausible that view now looks, in the light of developments in epistemology and other areas of enquiry. My conclusions will be somewhat pessimistic. The classical utilitarians are all, I shall suggest, intuitionists, but intuitionism runs into the problem of intransigent disagreement. Coherentist views may be plausible in themselves, but can be employed equally effectively by the opponents of utilitarianism in ethical theory. Nor do the approaches developed in the twentieth and twenty-first centuries look especially promising. I end with a plea for history.

1. What is utilitarianism?

Plato’s dialogue *Meno* begins abruptly, with the question to Socrates whether virtue can be taught. Socrates says that answering that question requires addressing a prior one: what *is* virtue? And, of course, that prior question gets Socrates and his interlocutors into all sorts of difficulties. Socrates’ claim about the priority of definition applies to utilitarianism as well. We can’t properly answer questions about, say, how demanding utilitarianism is, whether it

[1] B. Williams, ‘A Critique of Utilitarianism’, in J. Smart & B. Williams, *Utilitarianism: For and Against* (Cambridge, 1973), pp. 75-150, at 150.

requires unacceptable violations of rights, or indeed how plausible it is, without some account of what the theory actually is.

Fortunately, unlike ‘virtue’, a term in ordinary language, ‘utilitarianism’ is a piece of philosophical jargon, and to find out what it means we can consult the philosophers who invented it.[2] J.S. Mill claims to have brought the term ‘utilitarian’ into use, adapting it from a phrase in Galt’s *Annals of the Parish*.[3] He defined the view as follows:

> *UM*: Actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness.[4]

The notion of degrees of rightness and wrongness is, of course, odd, since these notions are bivalent. Any action is either right or wrong. But Mill’s meaning is clear enough. The right action is that which promotes happiness maximally, any action which fails to do that is wrong, and among wrong actions any action is morally worse to the extent that it fails to maximize happiness.

Mill almost certainly saw this as a version of what Jeremy Bentham had called ‘the principle of utility’, understood as a ‘standard of right and wrong’,

> which approves or disapproves of every action whatsoever according to the tendency it appears to have to augment or diminish the happiness of the party whose interest is

[2] It is true, of course, that the term has taken on a non-philosophical meaning, according to which to be a utilitarian is to rank usefulness over other values, such as moral goodness or beauty. This has led to widespread public misunderstanding of the basic tenets of the philosophical doctrine, which is especially unfortunate, since Mill almost certainly chose the term partly to avoid just such misconceptions.

[3] J.S. Mill, *Utilitarianism*, ed. R. Crisp (Oxford, 1998), 2.1n.

[4] Ibid., 2.2.2-4.

> in question: or, what is the same thing in other words, to promote or to oppose that happiness.[5]

The reference to ‘appearance’ here might be taken to introduce a ‘subjective’ conception of utilitarianism, according to which rightness and wrongness depend on, say, the value of the outcomes of various actions multiplied by the probabilities that would be ascribed to them by some rational subject. But that reference is dropped in 1.6, where Bentham speaks only of the tendency to augment happiness. And since Mill makes no reference to subjective considerations, the natural reading of him is also objective.

Bentham’s definition in *Introduction* 1.2 is neutral between rational egoism and impartial universalistic hedonism, and indeed other distributions of happiness, since ‘the party whose interest is in question’ may be an individual or a community. But it seems that what Bentham intends by reference to the individual is merely that the interest of any individual consists only in their happiness.[6] The principle of utility itself concerns the community – if we are to read the following as stating a necessary and sufficient condition of acceptance of the principle:

> A man may be said to be a partisan of the principle of utility, when the approbation or disapprobation he annexes to any action, or to any measure, is determined by and proportioned to the tendency which he conceives it to have to augment or to diminish the happiness of the community: or in other words, to its conformity or unconformity to the laws or dictates of utility.[7]

[5] J. Bentham, *An Introduction to the Principles of Morals and Legislation*, ed. J.H. Burns & H.L.A. Hart, intr. F. Rosen (Oxford, 1996), 1.1-2.
[6] *Introduction*, 1.5.
[7] Ibid., 1.9.

What is ‘the community’? Bentham sees it as nothing over and above the ‘individual persons’ that constitute it, and later he includes non-humans within the scope of benevolence.[8] Mill’s view is the same.[9]

UM, then, is in effect a restatement of the Benthamite position on utilitarianism. The third of the classical utilitarians, Henry Sidgwick, probably had both of his predecessors in mind when he defined utilitarianism as the view that:

> *US*: The conduct which, under any given circumstances, is objectively right, is that which will produce the greatest amount of happiness on the whole.[10]

Here we have an explicit reference to objective rightness, no commitment to degrees of rightness or wrongness, and clear recognition that the right action is that which maximizes happiness overall – that happiness is to be assessed impartially, with no special weight given to the happiness of the agent or those close to her. Since Sidgwick does not note any departure from the views of Bentham and Mill, I take it that he believed that *US* was essentially an interpretation of their own position which would have been acceptable to them. In that sense, then, *US* is perhaps *the* canonical statement of classical utilitarianism.

Utilitarianism is seen by the classical utilitarians as, at least in part, an answer to the question: ‘Which actions are right, and which wrong?’. The word ‘right’ here is ambiguous.[11] It could mean ‘*morally* right’ in a sense which allows that there may be, for example, self-interested reasons that are in some sense independent of morality and potentially in conflict with it. That, I suspect, is the sense in which Mill intended it in *Utilitarianism* 2.2, since he speaks of the view as one concerning ‘the foundation of morals’, and in the *System of Logic*

[8] Ibid., 5.10; see also 17.4n.b.

[9] *Utilitarianism*, 2.10.

[10] H. Sidgwick, *The Methods of Ethics* (7th edn., London, 1907), 4.1.1.2/411.

[11] And of course it has senses other than those I mention here, the most common perhaps being ‘correct’ (‘Yes, you’re holding the saw in the right way’).

he sees morality as just one ‘department’ of the art of life, the others being ‘prudence’ and ‘aesthetics’.[12] But consider the conclusion of the ‘proof’ in chapter 4:

> We have now, then, an answer to the question, of what sort of proof the principle of utility is susceptible. If the opinion which I have now stated is psychologically true -- if human nature is so constituted as to desire nothing which is not either a part of happiness or a means of happiness, we can have no other proof, and we require no other, that these are the only things desirable. If so, happiness is the sole end of human action, and the promotion of it the test by which to judge of all human conduct; from whence it necessarily follows that it must be the criterion of morality, since a part is included in the whole.[13]

Recall that, for Mill, morality is a part of the art of the life: it is that art which he is referring to here as ‘the whole’. On one interpretation of the passage here, Mill sees the art of life as involving only the notion that happiness is the only good. This might leave room for a ‘department’ of the art in which an agent is permitted or even rationally required to give special weight to her own happiness or that of those close to her. But this would allow for a conflict of principles within the art of life, and Mill will not allow that.[14] The ultimate principle of the art of the life *just is* the principle of utility in its broad sense. In other words, Mill seems to think that once I recognize and properly reflect upon the fact that the only good is happiness, I will grasp that my life as a whole should be directed towards maximizing happiness overall, and not my own happiness or that of some sub-set of persons. So, as far as utilitarianism is concerned, Mill see it as an answer to the broad question as well as the narrow, though the language used to answer the latter will be peculiarly moral.

[12] J.S. Mill, *A System of Logic*, 6.12.6, *Collected Works*, ed. J. Robson (Toronto, 1961-91), 8.949-50.

[13] *Utilitarianism*, 4.9.

[14] *System*, 6.12.7, *CW*, 8.951.

What about Bentham and Sidgwick? In a note of July 1822 attached to the term 'principle of utility', Bentham refers to the 'greatest happiness principle' as concerning 'the right and proper, and only right and proper and universally desirable, end of human action'.[15] So he might well have seen the question to which utilitarianism is an answer more broadly, as: 'Which actions do I have reason, overall, to do?'. And Sidgwick certainly took the question to be broad in this way. Ethics, he points out, is 'sometimes considered as an investigation of the true … rational precepts of Conduct', and he implies that we are interested in the principles that determine which conduct is ultimately reasonable.[16]

Should we begin with Mill's narrowly moral question, or the broader question, asked by all the classical utilitarians, about overall reasons for action? I suggest the latter, since this question is more fundamental or ultimate. As Williams says, the basic Socratic question 'How should one live?' means 'How has one most reason to live?', and 'no prior advantage is built into the question for one kind of reason over another'.[17] This fundamental question itself concerns which categories or types of reason there are, so it may turn out to be a mistake to assume the existence of a certain category at the start. Consider, for example, a rational egoist, who answers the broad question with the claim that the only reason anyone has is maximally to advance their own well-being. This egoist may just dismiss the moral question as irrelevant or misguided.

Sidgwick, like Bentham and Mill, couched his account of utilitarianism in terms of 'happiness', understanding that idea hedonistically. But here it probably makes best sense to

[15] *Introduction*, 1.1n. a.

[16] *Methods*, 1.1.2.1/2-3; 1.1.3.5/5-6.

[17] B. Williams, *Ethics and the Limits of Philosophy* (London: Fontana, 1985), p. 19. Williams suggests that the Socratic question is 'not immediate; it is not about what I should do now or next. It is about a manner of life' (ibid., p. 4). We might, on Sidgwick's behalf, seek to bridge the apparent gap here. We are indeed concerned about how our lives as a whole should go. But how they go depends on what we do here and now; and what we should do here and now may well depend on how our lives should go. And, here and now, what I should do here and now is the central ethical question (cf. Williams's earlier claim: 'Philosophers, not only utilitarian ones, repeatedly urge one to view the world *sub specie aeternitatis*, but for most human purposes that is not a good *species* to view it under', 'Critique', p. 118). Further, because of the direct link it establishes between utility-maximization and individual acts, act utilitarianism seems paradigmatic in comparison with forms of utilitarianism that focus on e.g. rules or motives.

depart from classical utilitarianism. Hedonism is a view about what is valuable; utilitarianism is a view about right action – that is, the actions we have most reason to perform. We should, then, understand utilitarianism more broadly as concerned with well-being rather than any particular conception of it. So now consider the following as a restatement of *US* as an answer to the broad question:

> *U*: Any agent has reason to act in such a way that the amount of well-being overall in the universe is maximized, and has no reason to act in any other way.

Note that *U* is consistent, for example, with a divine command theory, according to which:

> *DC1*: Any agent has ultimate reason to act in accordance with God's will, and has no ultimate reason to act in any other way.

and

> *DC2*: God wills that, and only that, the amount of well-being in the universe is maximized.

An ultimate reason is non-derivative, in the sense that, if I have an ultimate reason to φ, then the ultimate reason-giving property of my action is its being a φ-ing. So, according to *DC1*, the ultimate reason-giving property of acting in accordance with God's will is that so acting is acting in accordance with God's will.[18] As stated above, *U* is not restricted to

[18] The ultimate/non-ultimate distinction enables us to see also that there could be an Aristotelian version of utilitarianism, as well as a Kantian version, according to which, respectively:

ultimate reasons (which of course is why it is consistent with *DC*). So we might restate it as *Ultimate Utilitarianism*:

> *UU*: Any agent has ultimate reason to act in such a way that the amount of well-being overall in the universe is maximized, and has no ultimate reason to act in any other way.

The classical utilitarians, I believe, all accepted *UU*. Bentham says that pain and pleasure are 'sovereign masters', that '[i]t is for them alone to point out what we ought to do', and that the principle of utility is a 'foundation' and is not susceptible of proof because it 'is used to prove every thing else'.[19] The same goes for Mill. He speaks of happiness as an 'ultimate end', and his famous proof aims not at basing the utilitarian principle on some more fundamental principle but at making the case for the utilitarian principle's being itself fundamental.[20] In so far as he is a utilitarian at all, Sidgwick also accepts *UU*. This may seem doubtful, given his pessimistic suggestion in his important chapter on 'Philosophical Intuitionism' (3.13) that:

> There are certain absolute practical principles, the truth of which, when they are explicitly stated, is manifest; but they are of too abstract a nature, and too universal in their scope, to enable us to ascertain by immediate application of them what we ought

A1: Any agent has ultimate reason to act in such a way that her human good is maximally promoted, and no ultimate reason to act in any other way.
A2: The human good consists in maximizing the amount of well-being in the universe.

K1: Any agent has ultimate reason to act in accordance with the Categorical Imperative, and no ultimate reason to act in any other way.
K2: The Categorical Imperative requires agents to maximize the amount of well-being in the universe.

[19] *Introduction*, 1.2; 1.11; see 1.13.

[20] *Utilitarianism*, 1.5; ch. 4.

> to do in any particular case; particular duties have still to be determined by some other method.[21]

Later Sidgwick writes that the 'axiom of Rational Benevolence is … required as a rational basis for the Utilitarian system'.[22] That axiom is:

> *RB:* Each one is morally bound to regard the good of any other individual as much as his own, except in so far as he judges it to be less, when impartially viewed, or less certainly knowable or attainable by him.

If this principle is understood monistically, and not as one among several, it might appear to provide not just a basis for utilitarianism, but, if we take 'regard' as equivalent to 'promote', as something close to a statement of it.[23] Sidgwick, however, prefers to define utilitarianism as a version of hedonism, so for *RB* to constitute a statement of utilitarianism would require substituting 'happiness' for 'good'. And it is that substitution which is required for *RB* to become sufficiently concrete for it to constitute a practical method of ethics.[24] In other words, Sidgwick would be prepared to take *UU* as a true but partial statement of what we have ultimate reason to do.

Before closing this section, let me return once again to the notion of well-being. *UU* is consistent with the view that well-being is not the only value. Someone might accept, for example, that beauty is valuable in itself, but deny that this value itself provides any reason to act, except in so far as it promotes well-being through, say, the pleasure of its being

[21] *Methods*, 3.13.3.1/379.
[22] 3.13.5.1/387.
[23] For helpful discussion, see J. Schneewind, *Sidgwick's Ethics and Victorian Moral Philosophy* (Oxford, 1977), pp. 306-9.
[24] *Methods*, 3.13.5.4-5/388; 421n.1.

contemplated. This was clearly not the view of the classical utilitarians. Their hedonism is a version of welfarism:

W: Well-being is the only value.

I have suggested that we remove hedonism from our characterization of the canonical form of utilitarianism. What about welfarism? I suggest we leave it in, since a major part of the attraction of utilitarianism for many is the idea that, since nothing other than well-being is of value, it must be rational to promote it maximally. Further, if a utilitarian allows non-welfarist values, this may prove a hostage to fortune, since a deontologist, for example, may seek to introduce non-welfarist values of her own, such as respect for persons, desert, or even wrongness itself, on certain conceptions.

2. Should we believe utilitarianism?

2.1 The classical utilitarians, intuition, and coherence

UU is the canonical statement of utilitarianism. Should we believe it?

Bentham begins the *Introduction* with the claim that systems that seek to question the principle of utility 'deal in sounds instead of sense, in caprice instead of reason, in darkness instead of light', and as we have already seen he gives it a foundational role. That role is normative to the extent that even those who object to the principle unconsciously rest their argument on reasons themselves drawn from the principle.[25] Indeed Bentham goes so far as to claim that the word 'right' *means* 'utility-maximizing'.[26] It is tempting, then, to see Bentham as, broadly speaking, an intuitionist, in the sense that he believes that someone who

[25] *Introduction*, 1.13.

[26] Ibid., 1.10; 1.14 (10).

properly understands the utility principle, and the language in which it is couched, will grasp its correctness directly – that is, non-inferentially -- through reason as opposed to sentiment.[27]

Surely Mill cannot be said to be an intuitionist? He does admit that utilitarianism has to be accepted in the absence of 'what is commonly understood by proof' – that is, deductive proof. But he then goes on:

> We are not, however, to infer that its acceptance or rejection must depend on blind impulse, or arbitrary choice. There is a larger meaning of the word proof, in which this question is as amenable to it as any other of the disputed questions of philosophy. The subject is within the cognizance of the rational faculty; and neither does that faculty deal with it solely in the way of intuition. Considerations may be presented capable of determining the intellect either to give or withhold its assent to the doctrine; and this is equivalent to proof.[28]

By an appeal to intuition, Mill appears to have in mind an appeal to the utilitarian principle as something immediately obvious.[29] But if we allow intuition to include the rational grasp of that principle *after* reflection, then his position is similar to Bentham's. In other words, after reflectively grasping that pleasure is the only good, one will come rationally to accept utilitarianism.

In his core statement of the proof, Mill claims:

[27] See ibid., 1.14 (3).

[28] *Utilitarianism*, 1.5.

[29] See J.S. Mill, *Principles of Political Economy*, 8.5.2, *CW*, 2.124; *The Subjection of Women*, 3.8, *CW*, 21.305. Mill sometimes understands an appeal to intuition as an appeal to the feelings: see e.g.'Dr Whewell's Moral Philosophy', *CW* 10.193-4. But clearly he cannot mean that in this passage, where intuition is an activity of the rational faculty. The contrast is between intuitionistic 'blind impulse, or arbitrary choice' and the presentation of considerations to determine the intellect.

> No reason can be given why the general happiness is desirable, except that each person, so far as he believes it to be attainable, desires his own happiness. This, however, being a fact, 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: that each person's happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons. Happiness has made out its title as one of the ends of conduct, and consequently one of the criteria of morality.[30]

To this, Sidgwick objects:

> There being therefore no actual desire---so far as this reasoning goes---for the general happiness, the proposition that the general happiness is desirable cannot be in this way established: so that there is a gap in the expressed argument, which can, I think, only be filled by some such proposition as that which I have above tried to exhibit as the intuition of Rational Benevolence.[31]

But if the intuition of Rational Benevolence is, in effect, the principle of utility itself, it cannot fill any gap in Mill's argument. In fact, in so far as Mill's presentation of 'considerations capable of determining the intellect' is an argument (that is, a process of rational reflection rather than a strict deduction), that intuition is, as in Sidgwick, a conclusion.

Appeal to reflective intuition, then, is then central to the classical utilitarian argument. But it runs into serious problems. Consider the four conditions Sidgwick claims a 'significant

[30] *Utilitarianism*, 4.3.

[31] *Methods*, 3.13.5.4/388..

proposition, apparently self-evident' would have to meet fully were it to be established 'in the highest degree of certainty attainable'.[32]

1. Clarity. 'The terms of the proposition must be clear and precise.'

2. Reflectiveness. 'The self-evidence of the proposition must be ascertained by careful reflection.'

3. Consistency. 'The propositions accepted as self-evident must be mutually consistent.'

4. Non-disagreement. Sidgwick does not provide a succinct initial statement of this condition. He says:

> Since it is implied in the very notion of Truth that it is essentially the same for all minds, the denial by another of a proposition that I have affirmed has a tendency to impair my confidence in its validity. … And it will be easily seen that the absence of such disagreement must remain an indispensable negative condition of the certainty of our beliefs. For if I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgments necessarily reduces me temporarily to a state of neutrality. And though the total result in my mind is not exactly suspense of judgment, but an alternation and conflict between positive affirmation by one act of thought and the neutrality that is the result of another, it is obviously something very different from scientific certitude.

[32] Ibid., 3.11.2.2-7/338-42.

Sidgwick himself ran into difficulties with both the third and the fourth conditions, though he himself focused in particular on the third. Sidgwick found that when he reflected on utilitarianism, it appeared to be self-evident. But the same was true when he reflected upon rational egoism:

> It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental, and that consequently 'I' am concerned with the quality of my existence as an individual in a sense, fundamentally important, in which I am not concerned with the quality of the existence of other individuals: and this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual.[33]

The general thrust of Sidgwick's view here strikes me as highly plausible. Even if one does not go all the way to a completely partial, monistic egoistic position, the fact that some good or bad is going to be instantiated in one's own life and not in that of others seems normatively significant to the point that it threatens utilitarianism. Consider the following case, where the numbers represent hours of the most appalling agony:

Outcome 1

You	*Persons 2-1000*
50	1

[33] Ibid., CC 1.1/498.

Outcome 2

You	*Persons 2-1000*
1.05	1.05

The total number of hours of agony in outcome 1 is 1049, while in outcome 2 it is 1050. According to *UU*, your only reason here is to bring about outcome 1. It seems to me hard to deny that you have a reason to bring about outcome 2 because of the fact that in this outcome you will be experiencing much less agony.[34]

Now distinguish between *UU* as a monistic principle and the following non-monistic principle:

> *UUNM*: Any agent has ultimate reason to act in such a way that the amount of well-being overall in the universe is maximized.

UUNM is hard to deny. Consider a choice between the following two outcomes and assume all else is equal:

Outcome I

Person 1	*Person 2*
100	100

Outcome II

Person 1	*Person 2*
200	200

[34] It may also be claimed, of course, that fairness, equality, or justice speak in favour of outcome 2.

The property of the action that brings about outcome II that it will maximize well-being seems to count strongly in favour of it. That is, there is a strong intuitive case for *UUNM* (though an egoist will deny even this, of course). When it comes to *UU*, however, there will be significant disagreement independently of egoism, whether monistic or as one principle among others. For *UU* assserts that the *only* reason-giving property is an action's being utility-maximizing. So it will not only be egoists who deny *UU*, but all those who accept non-utilitarian reasons. And of course many non-egoists do. Not all these people have reflected especially carefully on their first-order normative views; but some have. Consider for example W.D. Ross:

> If, so far as I can see, I could bring equal amounts of good into being by fulfilling my promise and by helping some one to whom I had made no promise, I should not hesitate to regard the former as my duty…. [and] normally promise-keeping … should come before benevolence.[35]

Ross, then, appears committed to the following principle:

P: Any agent has ultimate reason to keep promises.

We have a stand-off, then, between the classical utilitarians and Ross.[36] The most sustained analysis and critique of a position such as Ross's – which we might call *reflective non-egoistic pluralism* (RNP), to distinguish it from apparently monistic theories such as

[35] W.D. Ross, *The Right and the Good* (Oxford, 1930), pp. 18-19.

[36] I shall not discuss virtue ethics, since it is most plausibly seen as a version of reflective (deontological or non-consequentialist) pluralism. See my 'A Third Method of Ethics?', *Philosophy and Phenomenological Research* (forthcoming 2012).

utilitarianism, from non-reflective common-sense morality, and from all forms of egoism – is of course Sidgwick's, in particular in the third book of the *Methods*. Sidgwick's main discussion of the principle of fidelity to promises is in sections 5-9 of the third chapter.

Sidgwick notes various qualifications that may have to be added to a principle such as *P*, interpreted in terms of an obligation on any agent who has made some promise to do what she and her promisee understood her to have undertaken to do. First, the promisee can waive the obligation. Second, a promise to perform an immoral act is not binding. Third, the bindingness of a promise may be affected by its dependence on a false statement by the promisee, though the limits here are a matter of dispute. Fourth, it may that the bindingness is affected also by a dramatic change in the circumstances, though again this is disputed. This becomes especially problematic when the promisee, perhaps because she has died, is no longer in communication with the promiser. Most would say that one should do what the promisee *would* have wished, but what this is very hard to say. And some will continue to insist that one should keep the original promise. Consider also cases in which the keeping of the promise will cause harm. If the harm is to the promiser herself, and is very great in comparison to the gain to the promisee, many will claim that the promise ought not to be kept. Consider also cases in which the promisee does not realize what harm will befall her from the promise's being kept. If the harm is extreme, no one will think the obligation persists. But then the question arises where to draw the line in less extreme cases. Sixth, problems arise in cases where misunderstandings have arisen over what exactly has been promised. We appeal here to customary language; but that may itself be unclear or changing, in which case it may be very difficult to settle the claims of each party. These difficulties of interpretation arise particularly in cases in which a person takes on some office by uttering some standard promise to the community. If there is disagreement about what that promise means, is the promiser allowed to choose any meaning she prefers, or should she act in accordance with the standard meaning? What if the formula in question is very old? Should it be interpreted as it was originally? And what is to be done if its meaning is in a process of alteration over time, and there is no consensus on which stage it has now reached?

Sidgwick's appeal to the existence of various disagreements here is certainly important and relevant to the credibility of any alleged principle of good faith. As we saw, he points out that disagreement between epistemic peers requires suspension of judgement. But any such disagreement is analogous to that over the principle of good faith itself between utilitarians and advocates of RNP. And his other highly acute and insightful remarks merely point out the various considerations which a conscientious moral agent may have to take into account in making her decision.

This very dependence on judgement, however, is unacceptable to the classical utilitarians, especially when that judgement involves the weighing of independent moral principles. From their point of view, a position such as Ross's is incomplete and unscientific. As Mill puts it:

> There are not only first principles of Knowledge, but first principles of Conduct. There must be some standard by which to determine the goodness or badness, absolute and comparative, of ends, or objects of desire. And whatever that standard is, there can be but one: for if there were several ultimate principles of conduct, the same conduct might be approved by one of those principles and condemned by another; and there would be needed some more general principle, as umpire between them.[37] (8.951)

[37] *System*, 6.12.7, *CW*, 8.951.

Mill, like Bentham and Sidgwick, is unwilling to allow a role for what Aristotle called *phronēsis*: a capacity to judge the weight of practically relevant factors, captured in a set of independent principles, on a case-by-case basis.

If we see their objection as primarily to pluralism, we should note the availability of alternative apparently monistic positions, such as, for example, a view according to which the right action is made right through its being such that the virtuous person would do it. But their antipathy to judgement in particular cases, informed rather than entirely guided by a set of principles, is unjustified, since it is not as if judgement plays no role in their own theories. First, it is in fact not obvious that utilitarianism is best described as monistic. It can be seen as a pluralistic view according to which we should maximize pleasure and minimize pain, judgement being required to balance the two. Even the various criteria for assessing pleasures and pains individually, such as intensity and duration, have to be judged against one another Second, judgement is required in accepting the theory in the first place (what is it if not judgement to have one's intellect determined by rational reflection on relevant considerations?). And, third, it will be required in applying the principle in particular cases, in deciding the value of different courses of action, their probabilities, and so on. Rawls's view that the debate between a systematic form of monism and pluralistic intuition cannot be decided *a priori* strikes me as more plausible:

> Now there is nothing intrinsically irrational about this [plural, reflective] intuitionist doctrine. Indeed, it may be true…. A refutation of intuitionism consists in presenting the sort of constructive criteria that are said not to exist.[38]

[38] J. Rawls, *A Theory of Justice* (rev. edn., Oxford, 1999), pp. 34-5.

But, of course, even if those constructive criteria can be presented in a form acceptable to both sides, we return to our stand-off – or rather our two stand-offs, between utilitarianism and egoism, and utilitarianism and reflective pluralism. Until the end of the nineteenth century, egoism was a powerful force in philosophy, and one might have expected Sidgwick's albeit reluctant acceptance of its plausibility to have promoted its reputation. But in the twentieth century its philosophical flame largely died out, though the embers were kept alive in other disciplines, such as economics and international relations, and in the implicit respect for egoism shown by its many philosophical critics. It seems to me nevertheless that egoism has not been refuted, especially as part of a pluralistic position, and it remains a serious challenge to utilitarians, even though it is not currently being pressed.

What about RNP? Given the intuitionist stand-off, it will be tempting for utilitarians to seek forms of argument other than an appeal to intuition. One natural direction is towards a form of coherentism, especially one based on alleged coherence between utilitarianism and common-sense morality (possibly in some reflective form). Bentham stated bluntly that all non-utilitarian principles are mistaken,[39] but Mill was ready to note utilitarian arguments in favour of veracity,[40] and Sidgwick took a good deal of time in books 3 and 4 of the *Methods* to explain the relation between utilitarianism on the ground that:

> [T]he Utilitarian argument cannot be fairly judged unless we take fully into account the cumulative force which it derives from the complex character of the coincidence between Utilitarianism and Common Sense.[41]

But coherentist arguments will result in another stand-off, since advocates of RNP, such as Ross, can plausibly claim that their own view provides at least as coherent an account

[39] *Introduction*, 2.1.
[40] *Utilitarianism*, 2.23.
[41] *Methods*, 4.3.1.3/425; see 3.13.1.2/373.

of common-sense moral intuitions as utilitarianism. Common sense provides no ruling on the central question of the plausibility of *P* – that is, on whether the obligation to keep a promise is grounded solely in utility, or stands on its own. Utilitarians will claim that their view provides solutions to many of the practical difficulties that emerge through conflicts between principles in pluralistic views and through application of those principles to particular cases. But here again it is open to the Rossian to appeal to the necessity for judgement in any reasonable form of theoretical ethics.

2.2 Modern alternatives

Might there be alternative non-intuitionistic ways of grounding utilitarianism in addition to attempts to find coherence between utilitarianism and the morality of common sense? The twentieth century offered various possibilities, but none has won wide support. Moore (surprisingly, given his own Open Question argument concerning goodness and pleasure) returned to the Benthamite assertion that '"right" does and can mean nothing but "cause of a good result," and is thus identical with "useful"' (PE, sect. 89).[42] Hare sought to derive utilitarianism from a sustained and brilliant study of the logic of the moral concepts (especially 'ought').[43] But he can be confronted with a form of Hume's is/ought dilemma. Either his argument relies on certain substantive, non-linguistic intuitions, which will be as controversial as *UU* itself, or his conclusion cannot itself be normative. J.J.C. Smart described his defence of utilitarianism as a presentation of 'Sidgwick in a modern dress', that is, as a statement of act utilitarianism without reliance on 'intellectual intuition'.[44] Rather, the axioms of utilitarianism are the 'expressions of our ultimate attitudes or feelings'. Sidgwick himself had already noted this line of argument:

[42] G.E. Moore, *Principia Ethica* (Cambridge, 1903), sect. 89.
[43] See esp. *Freedom and Reason* (Oxford, 1963); *Levels of Moral Thinking* (Oxford, 1981).
[44] J. Smart, 'An Outline of a System of Utilitarian Ethics', in *Utilitarianism: For and Against*, pp. 7-8.

> [I]t is not necessary, in the methodical investigation of right conduct, considered relatively to the end either of private or of general happiness, to assume that the end itself is determined or prescribed by reason: we only require to assume, in reasoning to cogent practical conclusions, that it is adopted as ultimate and paramount.[45]

Smart does not say whether he takes the attitudes to which he is appealing to be rational or non-rational. It may be, then, that Smart took himself to be appealing to the intellects of his readers rather than merely their tastes. If so, he may, like Mill, be offering considerations to determine the intellect. Or he may be seeking to persuade his readers to share his tastes; but that would leave him with no rational argument against those who do not. Finally, the attempt by Harsanyi and others to ground utilitarianism, without appeal to normative intuition, through reviving Adam Smith's notion of the impartial spectator, runs into problems similar to those faced by Hare.[46] The proposition that I have a reason to abide by the principles such a spectator would accept is itself ultimate, and so an intuition in the sense which matters. Further, this will anyway give us only non-ultimate utilitarianism, and leaves it open to pluralists to claim that an impartial spectator will be non-utilitarian.

Recently, Peter Singer, the most prominent contemporary utilitarian, has argued that evidence from evolutionary psychology and neuroscience may be adduced in support of utilitarianism against non-utilitarian views. He uses in particular the results of Joshua Greene's fMRI research on those considering trolley problems, according to which those prepared to cause harm to minimize it overall, by pushing a stranger off a bridge onto the trolley, show more activity in those areas of the brain associated with cognitive activity. This, Singer suggests, is exactly what we might expect when we reflect upon the evolutionary origins of morality in close, face-to-face interactions with others:

[45] *Methods*, 1.1.4.6/8.

[46] See e.g. J. Harsanyi, 'Morality and the Theory of Rational Behaviour', in A. Sen & B. Williams (ed.), *Utilitarianism and Beyond* (Cambridge, 1982), pp. 39-62.

> But what is the moral salience of the fact that I have killed someone in a way that was possible a million years ago, rather than in a way that became possible only two hundred years ago? I would answer: none.[47]

In other words, Singer's suggestion is that reflection on the origin of the non-utilitarian intuitions debunks them.[48] As Singer is aware, his argument, if it does throw doubt upon intuitions, does so only on those which consist in immediate and largely unreflective reactions to individual cases. If we carried out fMRI scans on Judith Thomson or Frances Kamm when they were articulating their anti-consequentialist views, we may well see broadly the same degree of largely cognitive activity as we might find in Singer himself or in those in Greene's experiment ready to push the stranger off the bridge.[49]

But what, anyway, about the Rossian principle *P*? The cases discussed by Greene involve 'close up and personal' harm to others. What would be the fMRI results if subjects were offered a choice between outcomes of equal value, which differ only in that in one outcome a promise is kept to someone who is dead or otherwise beyond harm as ordinarily understood? For an argument against principle *P* of the kind developed by Singer and Greene in the trolley case even to get off the ground, it would have to be the case that subjects choosing the outcome with promise-keeping would either show greater activity in the

[47] P. Singer, 'Ethics and Intuitions', *Journal of Ethics* 9 (2005), pp. 331-52, at 348.

[48] As Selim Berker points out ('The Normative Insignificance of Neuroscience', *Philosophy and Public Affairs* 37 (2009), pp. 293-329, at 319), consequentialist (or utilitarian) intuitions are also a product of evolution. But the Greene-Singer argument is that the evolutionary history of our emotional responses to close up and personal harm debunks any principle based on those responses. Berker continues: 'Sensing this sort of worry, Singer calls for us to engage in "the ambitious task of separating those moral judgments that we owe to our evolutionary basis and cultural history, from those that have a rational basis." However, this is clearly a false dichotomy' (p. 320). This is a somewhat uncharitable reading of Singer's claim: he means us to understand 'only' after 'owe'.

[49] Greene claims that deontologist theorizers are rationalizing their unreflective intuitions ('The Secret Joke of Kant's Soul', in W. Sinnott-Armstrong (ed.), *Moral Psychology 3: The Neuroscience of Morality* (Cambridge, MA, 2008), pp. 35-79, at 68; see Berker, 'Normative Insignificance', p. 315). Since there is clear selection value in benevolence, and benevolence itself will be emotionally detectable in particular cases in the same sorts of ways as disapproval of harming others, a *tu quoque* response is available to advocates of RNP.)

emotional areas of their brains than those neutral between the two outcomes, activity which could then be explained in terms of evolutionary theory, or similar activity, but activity such that it, and not the activity in the brains of the neutral group, could be explained in terms of evolutionary theory. My guess is the emotional activity in the brains of each group would be roughly the same, and that any explanation of one group's activity would be no more debunking than that of any explanation of that of the other group. There is research to be done here, but the chance of its outcome's supporting utilitarianism appears low.

What is perhaps more likely to bear fruit is further research on the history not of morality in general, but of *our* morality in particular. Just as moral philosophers have recently noted that neuroscience and evolutionary theory may be highly relevant to our understanding of ethics, so they now need to recognize (or to remember) that the same is true of our cultural history. When Bentham, Mill, and Sidgwick 'intuit' utilitarianism, or Ross intuits principle *P*, they are not doing so from some ahistorical, impartial standpoint. They are doing so *because* (of course only *in part* because) they have been brought up in a particular way in a particular culture. And again an understanding of that culture may prove to debunk certain moral intuitions.

As an example of what I have in mind, consider the remarkable case of supererogation. It is sometimes claimed that the fact that utilitarianism can make no room for genuinely supererogatory acts constitutes a serious problem for the the theory.[50] Certainly, the idea that one can go beyond duty, in a morally praiseworthy way, is now a basic assumption of the morality of common sense. But the idea was absent from the thought of Socrates, Plato, Aristotle, and all the pre-Christian Hellenistic philosophers, and appears only to have emerged as the Church Fathers in the second and third centuries began to develop the case for the monastic life on the basis of, in particular, the passage in *Matthew* 19: 16-21

[50] J. Urmson, 'Saints and Heroes', in A.I. Melden (ed.), *Essays in Moral Philosophy* (Seattle & Londo, 1958), pp. 198-216, at 206-7. For further discussion, see R. Crisp, 'Supererogation and Virtue', in M. Timmons (ed.), *Oxford Readings in Normative Ethics* (forthcoming 2013).

concerning the rich young man. The dominant view of supererogation now is that it is fundamental to morality. Reflection on its history at the very least puts that claim in some doubt. And it may be that promising is open to the same treatment. In modern times, the view that a promise is, other things equal, binding independently of the reason for its being made, emerged in the work of the natural lawyers in the sixteenth and seventeenth centuries.[51] Before that period, the English common law was based on the view that promises obligate the promiser only if they are made for some good reason, a view also found in France. As Patrick Atiyah puts it: 'These views … seem to have stemmed from the idea that a bare unilateral promise which has not yet been relied upon cannot cause any loss to the promisee. If it is not performed, the promisee is no worse off than he was before'. Atiyah admits that much of the reasoning in the arguments of Grotius and Pufendorf is 'extremely flabby', and goes on:

> It would hardly be worth serious consideration if it was not for the immense influence which the Natural Lawyers had, both with lawyers and philosophers; there is, too, a remarkable similarity between their whole approach, and that of many modern philosophers, notably, but not exclusively, intuitionists like Sir David Ross.[52]

The question for Ross, then, is whether, if it were the case that his intuition of principle *P* were the result of the influence of natural law theory on common-sense morality, and if that theory itself were based on poor reasoning, this would shake his confidence in *P*. Maybe it would not. But it is not entirely vain to hope that, if we understood ourselves and our history much better, there would be greater agreement in ethics, nor for the utilitarian to believe that there would be greater convergence on utilitarianism itself. To invert Gibbon: 'If

[51] P. Atiyah, *Promises, Morals, and Law* (Oxford, 1981), p. 9.
[52] Ibid., p. 11.

historians are not always philosophers, it were at least to be wished that all philosophers were historians'.[53]

Roger Crisp

St Anne's College, Oxford

Oxford Uehiro Centre for Practical Ethics

[53] E. Gibbon, *An Essay on the Study of Literature* (London, 1764), p. 107.