The Philosophical Quarterly Vol. 58, No. 231 *April 2008*
ISSN 0031–8094 doi: 10.1111/j.1467-9213.2007.518.x

VARIABLE *VERSUS* FIXED-RATE RULE-UTILITARIANISM

By Brad Hooker and Guy Fletcher

Fixed-rate versions of rule-consequentialism and rule-utilitarianism evaluate rules in terms of the expected net value of one particular level of social acceptance, but one far enough below 100% social acceptance to make salient the complexities created by partial compliance. Variable-rate versions of rule-consequentialism and rule-utilitarianism instead evaluate rules in terms of their expected net value at all different levels of social acceptance. Brad Hooker has advocated a fixed-rate version. Michael Ridge has argued that the variable-rate version is better. The debate continues here. Of particular interest is the difference between the implications of Hooker's and Ridge's rules about doing good for others.

Michael Ridge's argument that 'variable-rate rule-utilitarianism' is superior to Brad Hooker's version of rule-consequentialism is framed in terms of rule-*utilitarianism*, not rule-*consequentialism*.[1] But the difference between these is not relevant to Ridge's argument. Thus we here follow Ridge in framing the discussion in terms of rule-utilitarianism. Our paper briefly outlines Hooker's theory, replies to Ridge's arguments against it, and then challenges the plausibility of Ridge's theory.

I. FIXED-RATE RULE-UTILITARIANISM

Hooker's form of rule-utilitarianism holds that an act is wrong if it is forbidden by the code of rules whose internalization by 90% of people everywhere in each new generation has the greatest expected value. 'Each new generation' needs qualification so as not to include future generations which because of genetic engineering or some other technological breakthrough have different natures from ours.[2] The proposed cost/benefit assessment of alternative sets of rules is run on the basis that the rules are for internalization by 90% of everyone everywhere, i.e., 90% of each

[1] M. Ridge, 'Introducing Variable-Rate Rule-Utilitarianism', *The Philosophical Quarterly*, 56 (2006), pp. 242–53. Hooker's theory was presented in his *Ideal Code, Real World* (Oxford: Clarendon Press, 2000, henceforth *ICRW*).

[2] See R. Arneson, 'Sophisticated Rule Consequentialism: Some Simple Objections', *Philosophical Issues*, 15 (2005), pp. 235–51, at pp. 248–9; Hooker, 'Reply to Arneson and McIntyre', *Philosophical Issues*, 15 (2005), pp. 264–81, at p. 268.

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK, and 350 Main Street, Malden, MA 02148, USA

socio-economic group in each society. Hooker might have improved his formulation if he had used *approximately* 90%. In any case, he explicitly assumed that at least the basic moral rules are the same for everyone, rather than that different sets apply for different groups.

There are not only benefits but also costs associated with getting rules internalized. The greater the complexity and demandingness of rules, the greater the benefits of compliance. But also the greater the costs of internalization, since time, effort and psychological turmoil will be necessary costs of getting a rule about beneficence internalized by new generations. These costs recur with each new generation of infants, since each generation of infants is born with about the same predominant selfishness as previous generations of infants.

According to Hooker's formulation of rule-utilitarianism, the cost/benefit assessment of different possible requirements of beneficence focuses on the costs of getting a very demanding requirement of beneficence internalized by 90% of everyone, i.e., of the poor as well as of the rich. At some point, the benefits of widespread compliance with a more demanding requirement of beneficence would be outweighed by the costs of getting such a rule internalized by approximately 90% of everyone.

Hooker's argument for rule-utilitarianism begins with our various moral convictions. Other things being at least roughly equal, a moral theory is justified for us if it identifies a fundamental moral principle which both explains why our more specific considered moral convictions are correct and provides some impartial justification for them. Hooker contends that rule-utilitarianism does a better job than other theories of underwriting our confident reflective intuitions about which kinds of act are morally prohibited, which kinds are required, and which kinds are optional.

II. IS FIXED-RATE RULE-UTILITARIANISM CRIPPLINGLY ARBITRARY?

Ridge objects to assessing rules by considering the net benefit of only one 'fixed rate' of internalization. He well understands why it is important to assess rules by considering their internalization by less than 100% of everyone. There are many moral problems that could not exist in the ideal world where everyone completely internalizes a set of rules. For example, in an ideal world there need be no rules about how to deal with reprobates and amoralists. So a moral theory that considers internalization only by 100% of everyone would not formulate rules for how to deal with reprobates and amoralists. But any plausible moral theory must provide rules for dealing with such problems.

Thus we must consider the expected values of codes' internalization by something less than 100% of everyone. Hooker followed Richard Brandt in tying permissible action to what would be permitted by rules with the highest expected value if internalized by 90% of everyone. Ridge's first objection is that Brandt's and Hooker's version of rule-utilitarianism makes the fundamental principle of morality *arbitrary* in a certain way. After all, why go for 90%, instead of 99%, or 75%, or 25%?

Hooker (*ICRW*, p. 84) suggested 90% as a compromise between two considerations:

> On the one hand, we want a percentage close enough to 100% to hold on to the idea that moral rules are for acceptance *by the whole society of human beings*. On the other hand, we want a percentage far enough short of 100% to make salient the problems about partial compliance – such problems should not be thought of as incidental.

These two considerations have some force, even if they fail to compel the choice of 90%.

On the other hand, we think the objection that focusing on any one rate of internalization is arbitrary has very little force if the rules thus selected have no implausible implications. So do they have implausible implications?

III. FIXED-RATE RULE-UTILITARIANISM: ILL SUITED FOR OTHER LEVELS OF INTERNALIZATION?

Ridge's other objection is that Hooker's version of rule-utilitarianism might generate rules which are ill equipped to deal with less than 90% internalization. Ridge (pp. 245–6) offers an example about moral proselytizing in societies where some do not accept the relevant rules. If 90% of people accept the relevant rules, then moral proselytizing seems arrogant and overzealous. So a code appropriate for 90% internalization would forbid rather than require moral proselytizing. Then if the rules to follow are the ones whose internalization by approximately 90% has the highest expected value, proselytizing would be wrong. But, intuitively, if you are in fact in a society where a much smaller percentage (e.g., 60%) accept the rules, then your proselytizing might be a 'necessary evil' and thus obligatory (Ridge, p. 246).

Ridge thinks that Hooker's version of rule-utilitarianism cannot agree with the intuition that you would be morally required to proselytize in a society where only 60% have internalized rules with the greatest expected value.

In order to investigate this objection more deeply, we need to distinguish between two groups of rules. One group are rules which intuition holds that one should follow whether or not a high percentage of others accept good rules and thus have the moral views they should have. An example of such a rule is the rule that one should not kill innocent people. Intuitively, one is forbidden to kill innocent people whether or not other people's moral views are as they should be. The other group of rules are ones which intuition holds that one might not be morally required to follow when other people's moral views are not what they should be.

The point of distinguishing between these two groups of rules is that the first group is irrelevant to Ridge's objection to Hooker's version of rule-utilitarianism. Why is this group irrelevant? Ridge's objection is that Hooker's 'fixed-rate' version of rule-utilitarianism tells us to comply with the rules with the highest expected utility for 90% of the population to internalize, even when in fact something far less than 90% of the population have internalized this set. Intuition holds that one should follow the first group of rules whether or not a high percentage of others accept good rules. Suppose Hooker's version of rule-utilitarianism endorses this group of rules (because their internalization by approximately 90% of the population has the highest expected value). Then with respect to these rules, Ridge's objection is no objection at all, since, intuitively, we think these rules should be followed even when others do not accept them.

What, then, of the rules which intuition holds that one might *not* be morally required to follow when other people's moral views are not what they should be? If Hooker's version of rule-utilitarianism insists that one must follow these rules even when other people's moral views are not what they should be, Hooker's version of rule-utilitarianism is in trouble. But does Hooker's version of rule-utilitarianism have this implication?

Hooker contended that rule-utilitarianism leads to a plurality of rules, one of which requires us to 'do good for others generally up to some threshold of aggregate self-sacrifice'.[3] Unfortunately, when Hooker elaborated on this rule, he assimilated the aim of doing good for others with the aim of helping the needy (*ICRW*, p. 166). We shall formally merge the injunction to do good for others with the injunction to help those worse off. The merged rule is 'Do good for others, with some priority for those worse off, at least up to some threshold of aggregate personal sacrifice'.

Hooker also argued that rule-utilitarianism endorses a qualified rule that requires one to do what is necessary to avoid disaster, even if doing what is necessary to avoid disaster involves breaking other rules (*ICRW*, pp. 98–9, 135–6, 165ff.). The requirement to prevent disasters is qualified so as not to push one beyond significant aggregate self-sacrifice.

Perhaps the central case of helping others arises when one could help them avoid starvation or serious disease. Helping such people is doing good for others. It is also preventing disaster.

Apart from often agreeing in practice, the two rules 'Do good for others, with some priority for those worse off' and 'Prevent disaster' have other similarities. Obviously, neither requires one to sacrifice more in aggregate over one's whole life than some significant amount. (Admittedly, this is vague. But we cannot see how to avoid vagueness, in part because any more precise rule would fail to cover a wide enough range of situations.)

Another respect in which the two rules are similar is that each has a scalar component. The rule about doing good for others, especially for the worse off, up to some threshold of aggregate self-sacrifice, is meant to apply not only to cases where one is choosing between an act that would benefit others and an act that would not, but also to cases where one can choose between acts that would benefit others more and acts that would benefit them less. It counts in favour of an act that it would benefit others more. Likewise, the injunction to prevent disasters can be extended from cases where the choice is between an outcome that would be a disaster and one that would not, to cases where all available alternatives are disasters. In the latter cases, the rule about preventing disasters tells one to prevent the bigger disaster.

There is one more respect in which Hooker's rule-utilitarian rules about doing good for others and preventing disaster are similar. Neither of them limits how

[3] *ICRW*, p. 106. See also G. Cullity, 'Moral Character and the Iteration Problem', *Utilitas*, 7 (1995), pp. 279–89, and *The Demands of Affluence* (Oxford UP, 2004); T.M. Scanlon, *What We Owe to Each Other* (Harvard UP, 1998), p. 224.

much you can be required to do for others by pointing to how much each would need to do if everyone did his share (see *ICRW*, pp. 164–6). There are many kinds of situation in which you might be among people who are just as much in a position to help as you are but who are unwilling. Among alternative possible rules for how to behave in such cases, the rule whose internalization has the highest expected value is one requiring those who can to help, at least up to some aggregate level of self-sacrifice.

Returning to our test case, we are imagining that you live where only 60% of the population accept good rules. We are also taking as given that your proselytizing will (not just annoy others but) help improve their moral beliefs and behaviour. Presumably, if others' moral beliefs and behaviour improve in rule-utilitarian terms, net overall utility will increase. We assume that the increase in net overall utility includes benefits for at least some others.

So we are assuming that

(i) Your proselytizing would lead to a change in the moral beliefs and behaviour of others, and this change would benefit others enough to increase net good
(ii) You cannot identify something else you could do instead of proselytizing that would be likely to benefit others even more.

If (i) and (ii) are true, then whether or not the rule about doing good for others requires you to proselytize depends on whether it is the case that

(iii) Your proselytizing would not push you over the limit of required self-sacrifice.

Also salient might be whether the following propositions are true:

(iv) The increase in net good resulting from your proselytizing would be so large that failure to achieve that increase would constitute a disaster
(v) Nothing else you could do instead would prevent an even bigger disaster.

If (i)–(iii) are true, then even if (iv) and (v) are not true, the duty to do good for others requires you to proselytize. But if (iv) and (v) are true, then as long as your proselytizing would *not* push you over the limit of required self-sacrifice, the injunction to prevent disaster joins forces with the injunction to do good for others in calling for you to proselytize. Hence, far from being unable to explain why your proselytizing is morally right, Hooker's version of rule-utilitarianism can hold that two of its rules call for this action.

IV. RIDGE'S VARIABLE-RATE RULE-UTILITARIANISM

Ridge's variable-rate rule-utilitarianism holds that an act is morally required if it would be required by rules with the highest *average* expected utility for all different levels of social internalization (p. 248).

One might naturally assume that the way to find out which rules are right according to Ridge's theory is to do the mathematics for all the possible levels of compliance and then compute the average. Compute the expected utility for a set of rules on the assumption that 100% of everyone internalize the set (including the cost of getting people to internalize the set). Then compute the expected utility for this set on the assumption that 99% of everyone internalize the set (again, including the cost of getting people to internalize the set) and the same again for 97%, 96%, 95%, 94%, ... down to 1%. Then average these 100 numbers. Doing these 101 calculations gives the *average* expected utility for all different levels of internalization for this one set of rules. Now do the same exercise for every other set of rules you think might be serious competitors to the first set. The set with the highest average expected utility is the one determining what morality requires.

We have three objections to Ridge's account.

The first objection is that by using the *average* expected utility of *all* internalization rates, Ridge's theory can be skewed by an anomaly. Suppose we face a choice between two codes A and B which permit different acts. Suppose also that code A has high expected utility at *all* internalization rates, and code B has much lower expected utility at nearly every internalization rate, but a very high expected utility at one rate. Thus it might be that B turns out to offer the higher average expected utility, despite being inferior at all rates of internalization but one. A simplified version of this is represented in Fig. 1. Here Ridge's account implies that code B determines which acts are morally required, an unattractive implication unless we happen to know that there is 100% acceptance.

Internalization rate	Expected utility	
	Code A	Code B
100%	100	323
50%	100	−10
1%	100	−10
Average expected utility	100	101

Figure 1

The verdict which Ridge's variable-rate rule-utilitarianism gives would be especially troubling if the internalization rate responsible for skewing B's average expected utility is one that is impossible to attain. If a code has higher average expected utility only because of its value at a level of internalization that in fact cannot occur, how strange it would be to think that we should act in accordance with this code on account of its higher average expected utility.

Ridge might reply by modifying his theory so that it includes only *attainable* rates of internalization within the calculation of the expected utilities of codes. Alternatively, he could discount expected utilities for internalization rates from more distant possible worlds. These moves would remove some of the force of the objection. However, even if Ridge makes either of the moves just identified, it is still true that an unusually high expected utility for a code at one particular rate of internalization could cause a high average expected utility for the code even though it is poor at every other internalization rate.

Our second objection is that on Ridge's account, to compare just three different codes of rules would involve solving over 303 mathematical problems. Actually, if the codes have to be compared not just with integer percentages of acceptance, i.e., not only with 99%, 98%, 97%, etc., but with 99·9%, 99·8%, 99·7%, etc., then the number of mathematical problems rises to 3003. But even if codes are to be compared only with integer percentages of acceptance, requiring that such a comparison must be made is too epistemologically demanding.

Can Hooker's own theory do any better? An enormous amount of vagueness, imprecision and guesswork is involved in comparing codes. Furthermore, a number of different codes may have unsurpassed expected value, i.e., no other has greater expected value. According to Hooker's theory, when this happens, we should focus on these codes with unsurpassed expected value, and then pick whichever one of them is closest to conventional morality (*ICRW*, pp. 114–16). So Hooker's theory does not imply that the right rules to internalize and follow are unmanageably difficult to ascertain.

Ridge thinks the same is true of his variable-rate rule-utilitarianism. He suggests that without doing any mathematical calculations, we might be able to see that one possible moral code has higher expected utility at almost every possible rate of social acceptance than any alternative code has. Ridge suggests (p. 240) that this superior code secures not only all the same benefits that its rivals would secure but also some additional benefits as well, enough to outweigh any extra costs.

Ridge's reasoning seems to go as follows. Taking first the code with the highest expected utility for a population of whom approximately 90% accept that code, he seems to suppose that given how this code is selected, it does not need, and so would not have, rules for what to do in societies where much less than 90% accept the code. However, accepting this code might have low expected utility if one is in a population of which a low percentage accept the code.

So a code might instead contain rules which mention different levels of possible social internalization. Suppose this alternative code said 'When 100% accept good rules, do A, B, C, D; when 99% accept good rules, do B, C, D, E; when 98% accept good rules, do B, C, E, F; when 97% accept good rules, do B, C, F, G; …'. Because this fine-grained code explicitly caters for every different level of social internalization, compliance with it is likely to have better consequences when any percentage other than 90% accept good rules.

But against the increased net benefits of people's complying with more fine-grained rules, there are the added internalization costs of people's having to learn and remember so many fine-grained rules (fine-grained enough to distinguish between each different integer percentage of social internalization). The internalization costs for such fine-grained rules would be very high. Because of these costs, Ridge admits (p. 250) that the expected value of internalizing more coarse-grained rules might be higher:

> For example, perhaps the ideal code would include rules for conditions in which 'virtually everyone accepts the rules', where 'most people accept the rules', in which 'at least a majority accept the rules', … and where 'virtually nobody accepts the rules'.

To take stock so far, Ridge and Hooker disagree about which version of rule-utilitarianism's fundamental principle is best. They also disagree about the form of the rules derived from rule-utilitarianism's fundamental principle. The quotation above suggests that Ridge thinks the rules derived from his fundamental rule-utilitarian principle will explicitly distinguish between what to do when virtually everyone accepts the same rules as you, what to do when most people (but something short of virtually all) accept the same rules as you, what to do when a medium percentage accept the same rules as you, what to do when a significant minority of people accept the same rules as you, and what to do when virtually no one else accepts the same rules as you. In contrast, Hooker maintains that the only appropriate reference to others' internalization of and compliance with the same rules is one needed to discourage free riding.

The following table summarizes the contrast between Ridge and Hooker:

	Fundamental Principle	**Derived Rules**
Ridge	An act is wrong if it is forbidden by the rules which have at least as much average expected utility across all acceptance rates as any alternative rules have	• When virtually everyone accepts good rules, then do ABCDE • When most (but well short of all) people accept good rules, then do BCDEF • When a medium percentage of people accept good rules, then do CDEFG • When only a significant minority accepts good rules, do DEFGH • When virtually no one accepts good rules, do EFGHI • Whatever the level of acceptance, do XYZ[4]
Hooker	An act is wrong if it is forbidden by rules whose internalization by approximately 90% of everyone has at least as much expected utility as the same rate of internalization of any alternative rules	• Don't injure others, steal, or break promises • When deciding how to spend your own resources, give special weight to your family and friends • With some priority for the worse off, do good for others generally, at least until you reach a significant degree of aggregate self-sacrifice • Prevent disasters, at least until you reach a significant degree of aggregate self-sacrifice • To discourage free-riding, the above rules are qualified so as not to impose a requirement of kindness or self-denying restraint towards others who refuse to do the same for you

Ridge thinks that one of the advantages of his theory is that it leads to a number of rules which take a conditional form. We find Ridge's fundamental principle difficult to apply, and thus are not confident about what percentage of the rules it

[4] Ridge might accept that some rules do not need to be conditional on levels of acceptance. These rules might match some of the rules in Hooker's list of derived rules.

supports would take a conditional form. But suppose his derived rules about helping others do take a conditional form.

Now we can state our third objection to Ridge's principle. The rules about helping others which follow from Ridge's fundamental principle would presumably be something close to

- When virtually everyone accepts the same good rules, do at least a little to help worse off strangers
- When most (but well short of all) people accept relevant good rules, do at least somewhat more to help
- When a medium percentage of people accept relevant good rules, make at least a considerable contribution to help
- When virtually no one accepts relevant good rules, do *enormous* amounts to help, even if this involves *huge* personal sacrifices.

How would this last conditional rule operate in a world (arguably ours) where a very small percentage of people accept good (i.e., strong) rules about aiding others? The rule would make enormous demands. In fact, the demands would be so great that we think this rule too demanding to be intuitively plausible.

Here there is a contrast with the rule about aid derived from Hooker's fixed-rate rule-utilitarianism, i.e., the rule that agents are required to do good for others, with some priority for those worse off, at least up to some threshold of aggregate personal sacrifice. This rule is demanding, but not counter-intuitively so. Thus our conclusion is that Ridge's theory has counter-intuitive implications which Hooker's theory does not.[5]

University of Reading

[5] For helpful comments on earlier drafts we thank Robert Audi, Alan Carter, Tim Chappell, John Cottingham, Garrett Cullity, Jonathan Dancy, Daniel Dennis, Brian Feltham, Ian Folland, Keith Horton, Rob Lawlor, Paul McNamara, Tim Mulgan, Michael Ridge, David Sobel, Philip Stratton-Lake, Jussi Suikkanen, Christine Swanton and Fiona Woollard. Work on this paper was conducted while we had grants from the British Academy and the Arts and Humanities Research Council. We thank these funding bodies.