Measuring the Consequences of Rules

HOLLY M. SMITH

Pre-Publication Version

Rutgers, The State University of New Jersey

Rule-utilitarianism has recently enjoyed a resurgence of interest, aroused in part by the promise of contemporary versions of rule-utilitarianism to more successfully achieve the advantages classically claimed for rule utilitarian theories: to deliver prescriptions that match our pre-analytic moral intuitions about the moral status of individual acts; and to satisfy two significant claims – the claim that morality must serve to enhance human welfare, and the claim that morality must be universalizable.[1]

An intriguing new debate has now broken out about how best to formulate rule-utilitarianism – whether to evaluate moral codes in terms of the value of their consequences at a *fixed rate* (such as ninety per cent) of social acceptance (as Brad Hooker contends),[2] or to evaluate codes in terms of the value of their consequences *throughout the entire range of possible acceptance rates* (as Michael Ridge contends).[3] I shall introduce and argue that still a third formulation, optimum-rate rule-utilitarianism, achieves certain goals better than either Hook's fixed-rate rule-utilitarianism or Ridge's variable-rate rule-utilitarianism. But I shall also argue that none of these versions of rule-utilitarianism survive two criticisms that appear to broadly undermine rule-utilitarianism in any of its likely variants.

1. HOOKER'S FIXED-RATE RULE-UTILITARIANISM, AND RIDGE'S VARIABLE-RATE RULE-UTILITARIANISM

Early rule utilitarians tested whether or not moral code **C** constitutes an ideal set of moral rules by comparing the consequences of 100 per cent social *compliance* with **C** to the consequences of 100 per cent social compliance with rival sets of rules.[4] '100 per cent compliance' may be understood to mean 'no agent does what is prohibited by **C.**'

Subsequent rule utilitarians typically rejected this test in favor of assessing a moral code by reference to the consequences of the code if it were *accepted* by fewer than 100 per cent of the agents governed by the code. As Hooker describes this test, an agent who 'accepts' (which he calls 'internalizing') a code is understood to be someone who has dispositions to comply with the code, to encourage and form favorable attitudes towards others who comply, to feel guilt or shame when she breaks the code, to condemn and resent others' breaking it, etc., and furthermore believes that these dispositions are justified.[5] Such an agent, however, may not always do as the code prescribes – she may make mistakes in applying the code, or succumb to the temptation to promote her own interests or the interests of her loved ones rather than do what the code requires. Thus Hooker's formulation of rule-utilitarianism reads as follows:[6]

> An act is wrong if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of well-being (with some priority for the worst-off). The calculation of a code's expected value includes all costs of getting the code internalized. If in terms of expected value two or more codes are better than the rest but equal to one another, the one closest to conventional morality determines what acts are wrong.[7]

Hooker interprets 'overwhelming majority' to mean ninety per cent acceptance.[8] The move from a 100 per cent compliance test to a less-than-100 per cent acceptance test is substantially motivated by the need to devise a code that would include realistic

provisions for 'partial compliance' situations – situations in which not everyone does as morality requires. Unlike codes evaluated at 100 per cent compliance levels, an ideal code evaluated at a less-than-100 per cent acceptance rate will include such provisions, even though their inclusion makes it more burdensome to learn the code.[9]

Ridge objects to Hooker's fixed-rate rule-utilitarianism (hereafter, FRRU) primarily because it is *arbitrary* to fix ninety per cent acceptance as the key acceptance point. Why not eighty-eight per cent, or eighty-nine per cent, or ninety-one per cent or ninety-two per cent? What if only sixty per cent of the population accepts the moral code? Any successful code must be formulated to address issues that arise at any such level of acceptance. To choose a single exact acceptance point is arbitrary, and nothing so important about morality should be arbitrary.[10]

To avoid such arbitrariness, Ridge introduces 'variable-rate rule-utilitarianism' (hereafter, VRRU), which assesses a proposed moral code by estimating the consequences of that code *at all possible levels of social acceptance,* not just at some privileged fixed level of social acceptance such as ninety per cent. His own version of VRRU may be stated as follows:

> An act is morally right if it would be permitted (or required) by a moral code whose *average* expected value for all different levels of social acceptance is at least as high as that of any alternative code.[11]

Ridge's theory evaluates a code by calculating the average of the expected values of the consequences of a code's acceptance *at all possible rates of acceptance,* from 0 per cent through 100 per cent acceptance.

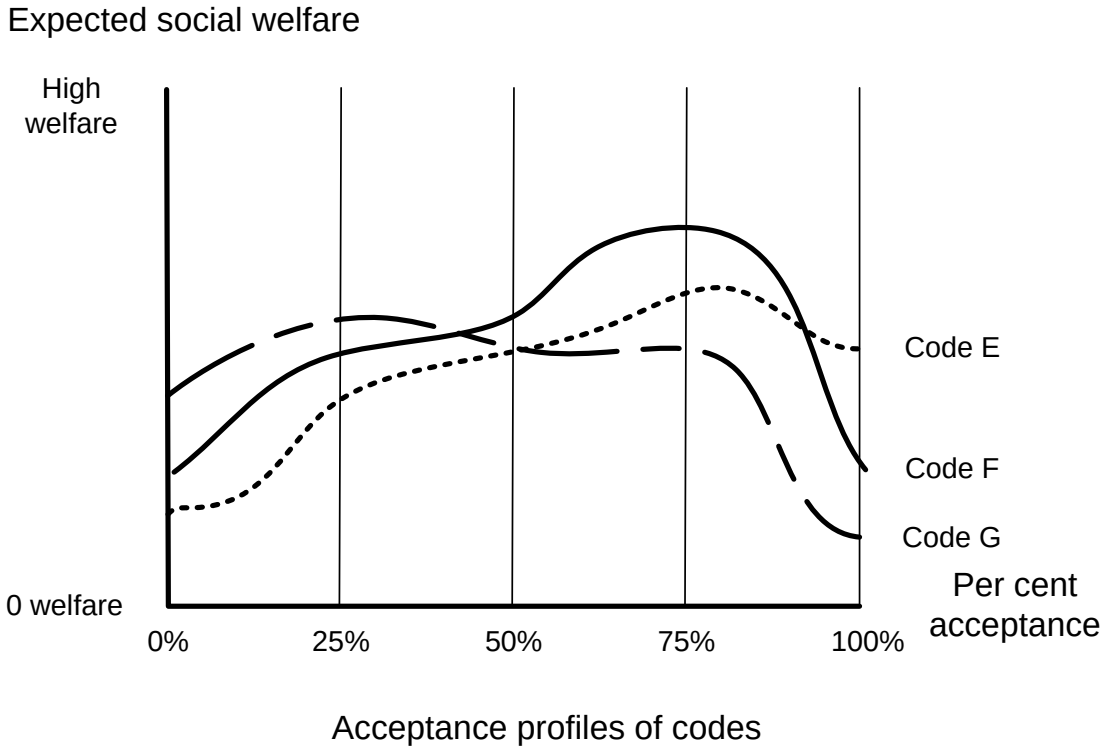
Ridge argues that his theory has all the advantages but none of the problems associated with FRRU,[12] and specifically that it avoids the charge of arbitrariness, since

it does not rely on identification of some specific level of social acceptance, but considers the consequences of a code across the whole spectrum of social acceptance rates.[13]

2. OPTIMUM-RATE RULE-UTILITARIANISM

Ridge's VRRU avoids the charge of arbitrariness by examining a moral code's acceptance-utility across every possible level of social acceptance. However, it is far from clear that this is the most intuitively attractive technique for avoiding arbitrariness. It is true that testing a code by its acceptance-utility at exactly ninety per cent acceptance seems arbitrary – but why test its acceptance-utility at every possible level, when in reality it must be accepted at some level or the other? Fortunately, there is another version of rule-utilitarianism utilizing an acceptance rate that is not arbitrary at all, and that seems to capture far better than either Hooker's or Ridge's versions the consequentialist spirit of rule-utilitarianism. To see this, recall that Hooker envisages that each generation would teach the ideal code to the next generation.[14] In selecting a code, the teaching generation will ask themselves 'Which moral code would produce the best consequences?' The consequences a given code produces are a function of two main factors: the content of the code, and the proportion of the population that accepts that code.[15] But the teaching generation can affect what proportion of the population accepts a given code through deft choice of teaching technique. For example, a harsher teaching technique may induce a larger proportion of the population to accept the code. Of course the techniques used to teach the code, and to secure its maintenance, themselves have an impact on social welfare, and Hooker correctly includes these costs as part of what has to be evaluated in appraising different codes. All this suggests that the teaching generation

faces a choice over what we can call ‘acceptance profiles’ of the different codes that might be taught, as represented in the following graph.[16]



Acceptance profiles of codes

An ‘acceptance profile’ of a code is the set of expected values it would produce at each of the possible levels of social acceptance (where the expected value of a code includes such items as the consequences of acts that comply with the code, the value of the code’s psychological ‘acceptance effects,’ the costs and benefits of teaching and maintaining it, etc.). Once we think in terms of code acceptance profiles, we can immediately see that – quite apart from the question of dealing with partial compliance -- it might be irrational

for the teaching generation to select a code in terms of the social welfare it would produce at 100 per cent acceptance. Although the expected social welfare produced by most normatively plausible codes rises as their acceptance levels rise, after a certain point the expected social welfare produced by many codes decreases as the increasingly burdensome cost of teaching and maintaining the more demanding or esoteric provisions of code outstrips the good procured by each degree of increased acceptance.[17] Agents will resist performing morally required actions when the personal cost is great, so that the effort to secure acceptance in these instances may outweigh the net good done by agents' acceptance of the code in such circumstances. (The acceptance levels profiled in this chart represent both how many agents accept the code in question, and how many occasions the agents accept the code in question as governing their choices *on that occasion.*) Thus the teaching generation would not necessarily select a code with the *aim of securing 100 per cent acceptance* of the code, since expected social welfare achieved by 100 per cent acceptance will often be less than the expected social welfare achievable at some lower level of acceptance. Moreover, the code that has the best consequences at 100 per cent acceptance may not have as good consequences at its optimum point as the optimum consequences achievable by some other code with worse consequences at 100 per cent acceptance. For example, compare Code E with Code F. At 100 per cent acceptance, code E would have better effects than Code F would at 100 per cent. Nonetheless, at the level of acceptance at which Code F would achieve its optimum effects – roughly seventy-five per cent acceptance -- the social value produced by Code F would outstrip both the social value produced by Code E at 100 per cent acceptance and the social value at the level of acceptance at which Code E would achieve its optimum

effects – roughly eighty per cent acceptance. The obvious thing for the teaching generation to do is to compare codes to see which code achieves the highest optimum acceptance level, where the optimum acceptance level for a code is the level of acceptance at which it achieves the greatest expected social good. The acceptance level that achieves this optimum will vary from code to code, and typically be less than 100 per cent acceptance.

Among the codes represented on this chart, Code F offers the highest optimum point. The teaching generation ought to select Code F to teach the new generation, and then aim, using the relevant teaching methods, to secure the acceptance level for Code F that would achieve this optimum point. Let us call the rule utilitarian theory that captures this idea 'optimum-rate rule-utilitarianism' (ORRU). It may be stated as follows:

> **ORRU:** An ideal code is the code whose optimum acceptance level is no lower than that of any alternative code.

Note that the ninety per cent acceptance level, the level on which Hooker focuses, is not necessarily the optimum acceptance level for any given code, or for the best code. Thus his theory irrationally advocates teaching and acting in accordance with a code (such as E) that may fall short of the optimum level of expected social utility achievable through some other available code. No teaching generation would make such a choice. Note also that once the teaching generation asks which code it should teach, there is no reason for it to consider the acceptance-utilities for the various codes at all the possible levels of acceptance. Contra Ridge's claim, they need only concentrate on which code will enable them to achieve the highest acceptance-utility, and teach that one.[18]

3. FRRU, VRRU, ORRU, AND THE PARTIAL COMPLIANCE PROBLEM

All versions of rule-utilitarianism must deal with the 'partial compliance' problem – the problem generated by the fact that, while a given moral code might produce excellent effects if everyone accepted or complied with it, it may produce extremely bad effects in a real-world situation in which there is only partial compliance or acceptance with the code. For example, Hooker describes a partial compliance case involving river pollution.[19] There, following Lyons,[20] he notes the distinction between cases involving *maximizing conditions* (where the agent needs to make the best of a generally good situation created by other agents' contributions to produce a public good) and cases involving *minimizing conditions* (where the agent needs to make the best of a generally bad situation created by other agents' failures to contribute to the production of a public good). In the pollution case, an important public good would be secured if almost everyone avoided polluting the river. If ninety per cent of the industries bordering the river dispose of their waste elsewhere, the river will be healthy. However, if fewer than ninety per cent dispose of their waste elsewhere, and more than ten per cent of the industries instead discharge waste into the river, the river will be dangerously polluted. On the other hand, suppose almost all the industries bordering the river discharge their waste into it so that the river is seriously polluted. If you are one of the industrialists, for you to join the few industrialists who avoid discharging waste into the river would be to impose a cost on yourself without producing any public good, since the river will be polluted whatever you do.[21]

How should a rule utilitarian deal with such cases? Ridge, like many rule utilitarians, argues that the best way to deal with the problem is to incorporate into the

moral code what I shall call 'conditionalized rules' prescribing or proscribing certain activities *conditional on* the level of a code's acceptance in society. He argues that the importance of avoiding heavy learning costs indicates that the best code would be one incorporating relatively few conditionalized rules for dealing with non-acceptance, stated with quite coarse-grained conditions, such as 'When most people accept the code, do X' and 'When less than half the people accept the code, do Y.'[22] For the pollution case, the relevant rules might be 'When most people accept the code, dispose of your waste elsewhere,' and 'When less than half the people accept the code, dispose of your waste in the river.'

How does Hooker deal with the problem of partial compliance? Hooker agrees that it would be unfair to require anyone to follow a burdensome rule for the sake of a public good when this rule is being ignored by most others, and agrees that rule-utilitarianism must provide a way to avoid this kind of unfairness.[23] He states that rule-consequentialism 'must not require agents to make sacrifices for others who are able to follow the same rule but won't'.[24] But exactly how does it avoid doing so?[25] The obvious suggestion is that his ideal code, like Ridge's, would include conditionalized rules permitting otherwise forbidden conduct when enough others are engaging in this conduct. Thus his FRRU might include such rules as the following:[26]

> **R***: If enough other people are failing to contribute to the production of a public good that your contribution would not secure its attainment, then you ought not to contribute if doing so is costly.

Thus both Ridge and Hooker can be interpreted as claiming that the best rule utilitarian solution to the partial compliance problem is to adopt conditionalized rules

specifying how an agent is to act as a function of how others are acting in the same situation. I shall assume that ORRU would adopt the same strategy.

4. A FATAL PROBLEM FOR FRRU, VRRU, AND ORRU

I will argue that Ridge's VRRU, Hooker's FRRU, and my ORRU – and indeed any version of rule-utilitarianism that invokes conditionalized rules to solve the partial compliance problem -- suffer from a fatal problem not noticed by either Hooker or Ridge. I shall begin by describing how the problem arises for Ridge's VRRU.

As we have seen, Ridge argues that the ideal code best handles partial compliance problems – at different levels of acceptance -- by including coarse-grained conditional rules of the form 'When virtually everyone accepts the code, do V,' or 'When less than half the people accept the code, do Y.'[27] The argument for a code incorporating such rules is that such a code would have higher average expected value than rival codes not incorporating conditionalized rules. Unfortunately this is false. A code incorporating conditionalized rules referring to the contributions of others towards producing some jointly-producible public good **may have no determinate expected value at all.**[28] In such cases the code cannot be judged to have a higher expected value than any rival code.

To see the problem, consider a slightly revised and more detailed version of the pollution case. In this version, three industrialists own factories bordering the river, you among them. If two or more of the industrialists discharge waste into the river, it will be severely polluted. But if one or none discharge waste, and the others burn their waste, the river will remain healthy, and no other environmental damage will be done (although each industrialist who burns factory waste pays a higher price for disposal than she would if she discharged waste into the river).

The relevant conditionalized rules for a code governing this case would be as follows:

Code C

R1: If two other industrialists are discharging waste, you ought to discharge.

R2: If one other industrialist is discharging waste, while the other is burning waste, you ought to burn your waste.

R3: If no other industrialist is discharging waste, you ought to discharge your waste.[29]

Suppose you are industrialist C. The following chart represents the possible choices that might confront you. For example, in Choice 1 the other two industrialists are both discharging their waste, and your choice is either to discharge (D) or to burn (B) your waste in this situation. The river will be polluted whichever choice you make.

	Choice 1				Choice 2				Choice 3				Choice 4			
Industries	D	B	D	B	D	B	D	B	D	B	D	B	D	B	D	B
A	x		x		x		x			x		x		x		x
B	x		x			x		x	x		x			x		x
C	**x**			**x**	**x**			**x**	**x**			**x**	**x**			**x**
Conseq's	Polluted river		Polluted river		Polluted river		Healthy river		Polluted river		Healthy river		Healthy river		Healthy river	

Figure 1

In these circumstances as industrialist C you have one possible occasion (Choice 1) on which your choice should be governed by rule R1; 2 possible occasions (Choices 2 and 3) on which your choice should be governed by rule R2; and 1 possible occasion (Choice 4) on which your choice should be governed by rule R3. Each of the other two industrialists has the same array of possible choices as seen from their perspectives.

Let us apply Ridge's theory to this situation. To take just the simplest question: what would the consequences be at 100 per cent acceptance - i.e. if all three industrialists accepted Code C? As soon as we ask this question, we can see that it has no determinate answer. For simplicity let us assume that each industrialist's accepting the code would result in his complying with it. If all three accepted and complied with Code C, there are two possible *abstract* patterns of discharging and burning waste, as shown in the following chart:

	Case 1		Case 2	
Industries	D	B	D	B
X	x			x
Y	x			x
Z	x		x	
Conseq's	Polluted river		Healthy river	

Figure 2

In Case 1, each industrialist appropriately accepts (and complies with) rule R1 of Code C. In Case 2, industrialists X and Y appropriately accept (and comply with) rule R2, while industrialist Z appropriately accepts (and complies with) rule R3. But these two patterns of action, predicated on universal acceptance of Code C, have very different consequences: in Case 1 the consequences (a polluted river) are very bad, while in Case 2 the consequences are very good (a healthy river, and one industrialist saves costs). Since either of these patterns of action (note Case 2 has several distinct realizations) involves universal acceptance of the code, but have very different consequences, it is indeterminate what the consequences would be of universal acceptance of the code.[30]

Ridge advocates assessing moral codes by calculating their average expected value, which is arrived at by ascertaining their expected values at various levels of acceptance. This only compounds the problem just seen. Consider a scenario in which there are 100 sets of three industrialists whose factories are located on the banks of 100 rivers, each industrialist facing the same waste disposal and polluting options as in the original story. At one level of acceptance, in which all 100 trios of industrialists accept and comply with Code C, the expected value of their doing so is indeterminate, since some trios may comply by all following rule R1, while other trios may comply by displaying appropriate patterns of following R2 and R3. But other levels of acceptance display the same indeterminacy. Suppose only ninety per cent of the industrialists accept Code C. Some trios of industrialists who are located together on one river will all accept and comply with Code C. But again, some of these trios may comply by all following R1, while other trios may comply by following R2 and R3. As we've seen, these two patterns of full compliance lead to very different consequences. So just by looking at these 'full acceptance' trios alone, without looking at what the non-accepters would do (and not even considering the complication that some industrialists who accept C may not successfully comply with its demands) we can see that there is no determinate answer to the question of what the consequences would be if ninety per cent of the industrialists accept C. And of course this will be true for every possible level of acceptance that involves at least some trios of joint code-accepters.

This shows that Ridge cannot plausibly hope to solve partial compliance problems through utilizing rules that incorporate conditions referring to what contributions other agents are making to the production of some jointly-producible public good (or bad). Of

course, it is unlikely that Ridge would advocate a version of VRRU that included rules such as R1 – R3, which are tailored specifically to one situation. But these rules reveal a problem that is more general, and crops up however such rules might be phrased, so long as they prescribe the same actions that R1 – R3 prescribe. Note especially that the advocate of VRRU cannot hope to avoid indeterminacy by rejecting overtly conditionalized rules and instead endorsing non-conditionalized rules such as 'Act so as to maximize social good in your circumstances' to guide agents in cases such as the pollution case. Since the recommendation of this rule for any agent depends on what the other agents are doing, this rule, too, counts as being universally followed in both Cases 1 and 2 by the agents in Figure 2, and must be seen as covertly conditionalized.

This problem is not unique to Ridge's VRRU. It is equally devastating for FRRU, and for my proposed ORRU, since they, too, attempt to deal with partial compliance cases by use of conditionalized rules. As I have argued, Hooker's best strategy for dealing with partial compliance cases is to adopt conditionalized rules similar to those utilized in Ridge's theory. Thus he should adopt (and may have considered himself as adopting) a set of rules that could be stated something as follows:

> **R***: If enough other people are failing to contribute to the production of a public good that your contribution would not secure its attainment, then you ought not to contribute if doing so is costly.
> **R****: If enough other people are contributing to the production of a public good that adding your contribution is necessary and sufficient for its attainment, then you ought to contribute.
> **R*****: If enough other people are contributing to the production of a public good to ensure its attainment, while your contribution would be personally costly, then you ought not to contribute.[31]

But these conditionalized rules are simply more general versions of rules R1 – R3, and they would be equally indeterminate for partial compliance problems arising at Hooker's fixed acceptance point of ninety per cent, as well as when applied to partial compliance problems (such as the pollution case) that arise at other contribution rates. My proposed ORRU would include similar rules, and fall prey to the same problem.

We have now found that Ridge's VRRU, Hooker's FRRU, and my ORRU use a technique for dealing with partial compliance cases that fatally undermines all three theories. Unless proponents of rule-utilitarianism can devise some successful strategy for dealing with partial compliance cases that avoids this flaw by eschewing conditionalized rules (or their equivalent), we have raised serious question whether any version of rule-utilitarianism can coherently give us the recommendations we need for partial compliance situations.[32] The use of conditionalized rules has sometimes been criticized on grounds that it leads to a rule-utilitarianism that is extensionally equivalent to act-utilitarianism. Couching rule-utilitarianism in terms of 'acceptance' rather than 'compliance' is often advocated on the ground that it avoids this extensional equivalence, but it does not avoid the problem of indeterminacy I have described here.

5. A SECOND FATAL PROBLEM FOR FRRU, VRRU, AND ORRU

Finally, I will describe a problem for these forms of rule-utilitarianism that has not previously been remarked. Consider again a rule's 'acceptance profile'. Previously I characterized the 'acceptance profile' of a code as the set of expected values it would produce at all the possible levels of social acceptance (where the expected value of a code includes such items as the consequences of acts in accord with the code, the value of the

code's 'acceptance effects,' the costs and benefits of teaching and maintaining it, etc.). However, this description of the components of a code's expected value fails to mention a critical element: the effects, for each given level of acceptance, of the associated level of *non-acceptance*. Thus if a code requiring agents to keep their promises has an acceptance level of sixty per cent, this means that sixty per cent of the agents accept an obligation to keep their promises (and/or accept this obligation on sixty per cent of the occasions on which they have made promises), and that forty per cent of the agents *don't* accept this obligation (and/or don't accept it on forty per cent of the occasions on which they have made promises). To calculate the overall expected consequences of sixty per cent acceptance of this code, we must know *both* the effects produced by those who accept it, *and* the effects produced by those who fail to accept it (let us call these the 'rejecters'). The implications of this fact seem to be overlooked by rule utilitarians in their discussions of what the consequences of a given code would be, since these discussions typically focus only on the effects of acceptance or compliance.

To know what the consequences would be of the actions of the rejecters, we must know what they would do instead of accepting the code. But what *would* they do? There are many different ways of not accepting a code. Suppose Code C* requires debtors to repay all their debts. There are many ways in which one could reject this code, and there are many different ways one might act in light of the fact one rejects C*. For example, one might reject this code and instead accept and follow a different code – but a code whose recommendations frequently coincide with those of C* in almost cases (for example, such a code might require debtors to repay all debts except those that would reduce the debtor to abject poverty). The acts of this type of C* code-rejecter would

frequently duplicate those of C* code-accepters. But there are also many ways of rejecting C* that are likely to lead to very different actions from those performed by accepters of Code C*. For example, one might reject C* but accept a code whose recommendations often diverge from those of C*, but still require certain efforts to repay debts (such a code might permit partial or delayed repayment of debts when the debtor is hard-pressed for funds). Or one might reject C* but accept a code whose recommendations diverge from C* in being *more* demanding (such a code, pinning honor on reciprocal generosity, might dictate that debts should be repaid at double their amount). Or one might reject C* but accept a code that permits avoiding repayment of debts by killing those to whom one inconveniently owes money. Or one might reject C*, and indeed reject all codes, in favor of acting purely out of whim or self-interest. The possible ways of rejecting a given code are legion, and they can lead to very different actions, even among agents facing the same situation. This is not merely a logical possibility; it seems *likely* that different rejecters will reject any given code in many different ways.

Each of these ways of 'not accepting' the code would have different effects on social welfare. The consequences of a sixty per cent acceptance rate for a given code depend heavily on what the code rejecters would do, and it is not at all clear that we have any reliable way to determine what this would be – especially for cases in the further future.[33]

Of course a similar point could be raised about the multitude of different ways of *accepting* a code (quite apart from the fact that one can accept a code but actually fail to comply). A debtor could repay a $100 debt by handing the lender a $100 bill, or by

handing the lender ten $10 bills, or by writing a check, or by using PayPal, or by sending a money order through the mail, etc. In some cases these different ways of accepting the code would have different consequences (perhaps the debtor's handing the money in person to the lender would lead to further beneficial interactions between them, whereas sending a money order would end their interactions). However, insofar as a given code specifies the agent's duty in fairly concrete terms, the room for variance in carrying out the code's prescriptions may be somewhat less problematic than the room for variance in ways in which agents can fail to accept the code.[34]

Unless we have a way of determining what the code-rejecters would do, we can't calculate the expected value of a given code for any acceptance level below 100 per cent. And as an empirical matter, answering this question seems like an insurmountable hurdle. A proponent of rule-utilitarianism might attempt to solve this problem by simply *stipulating* what the rejecters would do, rather than trying to *ascertain* what they would do by determining the truth of the relevant counterfactuals. For example, it might be stipulated that the rejecters would follow self-interest rather than accept the code. Or it might be stipulated that the rejecters would obey the prevailing morality rather than the code under consideration. But how is the theorist to non-arbitrarily pick which stipulation to make? I cannot see any reasoned way to do this.

An advocate of rule-utilitarianism might hope that it makes no difference which baseline manner of non-acceptance we choose for comparing the consequences of each code at any given level of acceptance. If this were true, some Code C* could be shown to be superior to a rival Code C** whether we stipulate that all the rejecters pursue their self-interest, or stipulate that all the rejecters follow the prevailing moral code, or

whatever. Hence we could select this baseline arbitrarily without affecting our determination of which code is best.[35] This hope might seem to be borne out by the following chart, in which an agent's available acts (A, B, C, etc.) are shown arrayed against columns designating which act would be chosen by which code (e.g. Code C* prescribes act B), which act would maximize self-interest (act A), which act is prescribed by the prevailing morality (act C), and the net social value that would be produced by each act.

Act	Code C*	Code C**	Self interested act	Prevailing morality	Net social value
A			X		90
B	X				100
C				X	10
D		X			80

Figure 3

Suppose we arbitrarily stipulate that the 'baseline' act that would be performed by a code rejecter is **the self-interested act** (that is, any agent who fails to accept the code performs the self-interested act). Thus, consider two agents who have precisely the same array of options shown in Figure 3, and let us compare the consequences of a fifty per cent acceptance rate with Code C* and alternatively with Code C**. The comparison between **(a)** the consequences of a fifty per cent acceptance rate for Code C* (one compliant act B with a net social value of 100, plus one non-compliant act A with a net social value of ninety = 190) and **(b)** the consequences of a fifty per cent acceptance rate for Code C** (one compliant act D with a net social value of eighty plus one non-compliant act A with a net social value of ninety = 170) would be as follows:

Act	Code C*	Code C**	Self interested act	Net social value	Net social value of 50% acceptance rate Code C*	Net social value 50% acceptance rate Code C**
A			X	90	190	170
B	X			100		
D		X		80		

Figure 4

Suppose instead we arbitrarily stipulate that the 'baseline' for acts that would be performed by a code rejecter is **the act prescribed by the prevailing morality**. Then (again for a group of two agents) the comparison between **(a)** the consequences of a fifty per cent acceptance rate for Code C* (one compliant act B with a net social value of 100, plus one non-compliant act C with a net social value of ten = 110) and **(b)** the consequences of a fifty per cent acceptance rate for Code C** (one compliant act D with a net social value of eighty, plus one non-compliant act C with a net social value of ten = ninety) would be as follows:

Act	Code C*	Code C**	Prevailing morality	Net social value	Net social value 50% acceptance rate Code C*	Net social value 50% acceptance rate Code C**
B	X			100	110	90
C			X	10		
D		X		80		

Figure 5

Whichever baseline we use as the act a rejecter would perform, Code C* achieves a higher net social utility than Code C**. So it appears that it makes no difference which baseline we select as the act the code rejecter would perform, and hence that we need not worry about arbitrarily stipulating what act to utilize as the one that a code-rejecter would perform.

Unfortunately, this appearance is misleading, as are these charts. The argument goes wrong in that it does not take into account *the value generated by the acceptance of some moral code in the society.* General acceptance of a moral code would (for example) generate psychological costs for violating the code – social disapproval from one's peers, and in some cases feelings of guilt in the code rejecter – which must be included in the net social value. If we add these in, it is possible to describe cases in which the highest net social utility is produced by C* if the baseline comparison for rejecter acts is the self-interested act, but is produced instead by C** if the baseline comparison for rejecter acts is the act prescribed by the prevailing morality.[36]

To see this, let's assume that social disapproval experienced by the agent of a non-compliant act has a disvalue of minus twenty-five, while the feeling of guilt for not accepting and following the code on a given occasion has a disvalue of minus five. Let's

also change the case so that the act prescribed by Code C** is the *very same act* as that prescribed by self-interest, i.e. act A. We then get the following comparison.

As before, we begin by stipulating that the 'baseline' for acts that would be performed by a code rejecter is the **self-interested act**. Thus (for a group of two agents) the comparison between **(a)** the consequences of a fifty per cent acceptance rate for Code C* (one compliant act B with a net social value of 100, plus one non-compliant act A with a net social value of ninety, plus a total of minus thirty from the costs of social disapproval and guilt = 160) and **(b)** the consequences of a fifty per cent acceptance rate for Code C** (one compliant act A with a net social value of ninety plus one non-compliant act A with a net social value of ninety, plus a total of minus five for feelings of guilt = 175) would be as represented in the following chart. Notice that since the act prescribed by Code C** is the very same act as the self-interested act, the rejecting agent will perform exactly the same act as he would if he accepted and followed Code C**. Although he may feel guilt (in acting for the wrong reason), he will not experience social disapproval, since no one will know that he was not 'following' Code C**.

Act	Code C*	Code C**	Self interested act	Net social value	Net social value 50% acceptance rate Code C*	Net social value 50% acceptance rate Code C**
A		X	X	90	160	175
B	X			100		

Figure 6 (baseline = self-interested act)

Suppose instead we stipulate that the 'baseline' for acts that would be performed by a code rejecter is the **act prescribed by the prevailing morality**. Then (again for a group of two agents) the comparison between **(a)** the consequences of a fifty per cent

acceptance rate for Code C* (one act B with a net social value of 100, plus one act C with a net social value of ten, plus a total of minus thirty from the costs of social disapproval and guilt = eighty) and **(b)** the consequences of a fifty per cent acceptance rate for Code C** (one act A with a net social value of ninety, plus one act C with a net social value of ten, plus a total of minus thirty from the costs of social disapproval and guilt = seventy) would be as follows:

Act	Code C*	Code C**	Prevailing morality	Net social value	Net social value 50% acceptance rate Code C*	Net social value 50% acceptance rate Code C**
A		X		90	80	70
B	X			100		
C			X	10		

Figure 7 (baseline = act prescribed by prevailing morality)

In these comparisons, Code C** produces higher social value when the baseline is the self-interested act, whereas Code C* produces higher social value when the baseline is the act prescribed by the prevailing morality.

Thus it *can* make a difference in comparing candidate codes what baseline we utilize for the acts that code-rejecting agents would perform. We cannot, then, solve this problem, as the rule utilitarian may have hoped, simply by arbitrarily stipulating a baseline for non-accepting acts. Since our stipulation makes a difference to which code has the best consequences, we need a compelling reason to make one stipulation or another – but there isn't any compelling reason to choose any given baseline.

My conclusion is that all forms of rule-utilitarianism are fatally flawed unless some solution can be found to the problem of determining, or specifying, what the code-

rejecters would do instead of accepting the code under consideration. For if we cannot determine what the rejecters would do, we cannot say what the expected consequences of any given level (short of 100 per cent) of acceptance of a code would be – and hence we cannot compare the consequences of rival codes in order to determine which code is ideal.

6. CONCLUSION

I have described three different approaches to measuring the consequences of rules for purposes of formulating rule-utilitarianism: Hooker's fixed-rate rule-utilitarianism (which evaluates rules by their consequences at a fixed rate of acceptance, such as ninety per cent), Ridge's variable-rate rule-utilitarianism (which evaluates rules by their consequences at every possible rate of acceptance), and my own suggestion, optimum-rate rule-utilitarianism (which evaluates rules by their consequences at their optimum acceptance rate). I have then argued that all three of these versions of rule-utilitarianism fall prey to the same two fatal objections. The first objection is that for all three theories, the preferred method for dealing with partial compliance cases invokes the use of conditionalized rules (according to which the best act for the agent is conditional on what other agents do), but these rules leave the expected social consequences of any code incorporating them completely indeterminate. The second fatal objection is that the evaluation of a code, as judged according to the expected social consequences of its acceptance (either at a single privileged rate lower than 100 per cent, or at all possible rates) will turn partly on what actions are performed by those agents who reject it. Since there are so many different ways to reject a code, what these code rejecters *would do*

seems simply indeterminate. Moreover, there is no non-discriminatory way of arbitrarily stipulating what baseline rejection act the rejecters would perform. Hence again there is no way of measuring and comparing what the consequences of rival moral codes would be.

Unless an improved version of rule-utilitarianism is found that evades these two objections, we must be pessimistic about the ultimate fate of this type of theory. [37]

hsmith@philosophy.rutgers.edu

[1] See discussion in Tim Mulgan, *Future People* (Oxford, 2006), pp. 130-33.

[2] Brad Hooker, *Ideal Code, Real World* (Oxford, 2000).

[3] Michael Ridge, 'How to Be a Rule-Utilitarian: Introducing Variable-Rate Rule-Utilitarianism,' *The Philosophical Quarterly* 56, No. 223 (2006), pp. 242-53. For Hooker's response, see Brad Hooker and Guy Fletcher, 'Variable *versus* Fixed-Rate Rule-Utilitarianism,' *The Philosophical Quarterly* 58 (2008), pp. 344-52. See also Richard Arneson, 'Sophisticated Rule Consequentialism: Some Simple Objections,' *Philosophical Issues* 15 *Normativity* (2005), pp. 235-51, and Brad Hooker, 'Reply to Arneson and McIntyre,' *Philosophical Issues* 15 *Normativity* (2005), pp. 264-81.

[4] Brad Hooker prefers the label 'rule consequentialism,' while Michael Ridge prefers 'rule utilitarianism.' I shall follow Ridge's usage.

[5]Hooker, *Ideal Code*, p. 76. Here Hooker follows a number of earlier theorists.

[6] See note 4.

[7] Hooker, *Ideal Code*, p. 32. Hooker later points out that the costs of maintaining and reinforcing the code must also be included (Hooker, *Ideal Code*, p. 79).

[8] Hooker, *Ideal Code*, pp. 83-4.

[9] See Hooker, *Ideal Code*, pp. 80-5.

[10] Ridge, 'How to Be,' pp. 244-45.

[11] Ridge, 'How to Be,' p. 248; see also the statement of the theory in Hooker and Fletcher, 'Variable *versus* Fixed-Rate,' p. 348. I have slightly reformulated Ridge's statement to allow for acts that are permitted (but not required) by the code.

[12] Ridge, 'How to Be,' p. 253.

[13] Ridge, 'How to Be,' p. 248-49.

[14] Hooker, *Ideal Code*, p. 32; Brad Hooker, 'Reply,' pp. 267-9.

[15] Of course, the circumstances in which each action would be performed play a major role as well.

[16] It is possible that some codes have 'feasibility' gaps in their possible acceptance levels that are not represented in this graph.

[17] This is not necessarily an argument for reducing the demands of even the best code, since the existence of these demands, even if unmet in practice, may inspire agents to try harder and comply more often than they would if the code made no such demands.

[18] Because of its focus on the *highest achievable social value*, ORRU appears to best represent the consequentialist spirit of rule-utilitarianism even if we don't adopt the perspective of the 'teaching generation.'

[19] Hooker, *Ideal Code*, pp. 124-25.

[20] Hooker, *Ideal Code*, pp. 124-25; David Lyons, *Forms and Limits of Utilitarianism* (Oxford, 1965), pp. 128-31.

[21] Hooker, *Ideal Code*, p. 123.

[22] Ridge, 'How to Be,' pp. 249-50.

[23] Hooker, *Ideal Code*, p. 124.

[24] Hooker, *Ideal Code*, p. 125.

[25] Some of Hooker's remarks suggest that he might invoke his 'Prevent disasters' rule as a way of dealing with partial compliance cases (see Hooker, *Ideal Code*, p. 98). However, this is not a satisfactory general solution. Hooker is vague on exactly what counts as a 'disaster,' but he stipulates that a disaster must involve 'large losses in aggregate well-being,' although 'there are limits to how much self-sacrifice can be demanded in the name of this rule' (Hooker, *Ideal Code*, p. 121). If the public 'bads' in my cases are deemed to be 'disasters,' we simply need to substitute cases in which the bad effects are smaller scale. As he sees, it is not open to Hooker to define a 'disaster' as *any* consequence which has net negative utility (Hooker, *Ideal Code*, p. 98).

[26] Hooker's FRRU evaluates all rules by the ninety per cent acceptance test. There will be some minimizing-condition situations in which a public good will only be produced if *more than* ninety per cent of the population – say, ninety-five per cent -- contribute to its production, although 100 per cent contribution is not necessary. Hooker needs a solution to such situations, and conditionalized rules seem to be his best option. Hooker seems to have no problem with the general idea of codes incorporating conditions when the conditions refer to such facts as the degree of intelligence of the agent, or whether or

not the agent is a parent. His discussion suggests openness to the conditionalized rules solution in partial compliance cases.

Of course many situations involving minimizing conditions are ones in which a threshold contribution level somewhere *below* ninety per cent -- say, eighty per cent contribution -- is required to produce the public good. Perhaps Hooker's best strategy for dealing with these cases is to argue that his theory endorses a rule such as R* for situations involving a greater than ninety per cent threshold, and that this rule can be stated with enough generality that it covers all the cases at the lower thresholds as well. Even though his ninety per cent acceptance level doesn't permit him to argue *directly* for the necessity of including rules that cover the lower threshold cases, there is no obvious reason why the ideal theory cannot include rules that are necessary at the ninety per cent level and also work well at lower acceptance levels. The only kinds of cases for which this strategy would not succeed would be cases in which the nature of the situation requires that the rule's condition must overtly specify that contribution levels are less than ninety per cent. Such cases would probably be few in number, so it appears that Hooker may avoid most of the difficulties arising in partial compliance cases by using the strategy I have just described.

[27] Ridge, 'How to Be,' p. 250. Ridge states these conditionalized rules in terms of people's acceptance of *rules* rather than acceptance of *codes*. I have simplified the discussion by restating the suggestion in terms of codes.

[28] I originally argued for this point in 'David Lyons on Utilitarian Generalization,' *Philosophical Studies* 26 (1974), pp. 77-94. The point was later stated in a broader theoretical context by Donald Regan in *Utilitarianism and Co-operation* (Oxford, 1980), p. 87.

[29] The situation is best understood as one in which you cannot form an agreement with the other industrialists regarding your conduct.

An alternate, and perhaps deontologically more attractive, version of rules R1and R3 would permit you to either burn or discharge your waste. But in the cases covered by these rules, it would reduce net social welfare for you to burn the waste, since doing so costs you more money, and doesn't affect whether or not the river would be polluted. A version of this case incorporating such permissive rules would, in any event, have the same problem that I describe in the text for rules R1 and R3.

[30] There are two distinct interpretations of this charge. One is that such conditionals as 'If all three industrialists accepted code C, the river would not be polluted' have no determinate truth value. The other interpretation is that such conditionals may have a truth value, but that (at least in most cases) the different patterns of action that count as 'accepting the code' mean that we are unable to ascertain what their truth value is, so we cannot ascertain which moral code would be best. Certainly the latter is true for complex codes purporting to govern the conduct of an indefinite number of agents facing partial compliance situations (even though we might feel we could ascertain what some particular trio of industrialists would actually do if they accepted Code C). I shall not attempt to mediate between these two interpretations.

[31] See note 29 on versions of these rules including permissions to contribute even though doing so is costly. Hooker (Hooker, *Ideal Code*, pp. 124-25) approvingly cites Brandt's dismissal of any consequentialist case for permitting an agent to fail to contribute to some public good when enough others are already contributing (i.e. in a maximizing case). Brandt dismisses such rules on grounds that it would be all too easy for most people to believe that a sufficient number were already contributing (Richard Brandt, 'Some Merits of One Form of Rule-Utilitarianism,' *University of Colorado Studies in Philosophy* (1967): n. 15, as cited in Hooker, *Ideal Code*, pp. 124-25). Of course sometimes this is so, but on other occasions it may be crystal clear that one's own contribution is not needed. A genuine consequentialist solution would prescribe not contributing when doing so would maximize social welfare.

[32] The elaborate theory labeled 'co-operative utilitarianism' advanced by Donald Regan in *Utilitarianism and Co-operation* may successfully avoid this problem.

[33] The complexity of this determination is compounded by the fact that the different choices selected by each type of code-rejecter at some base time (say, now) will quickly ramify into a wholly different set of opportunities and choices faced by them in the future that are not faced by the code accepters or by agents who reject the code in favor of different options. The complexity is even further compounded by the fact that we must consider what the code-rejecters would do in a context in which a significantly different code is imagined as being in force, as compared with the actual world.

See note 30 on the question of whether the indeterminacy in question is an indeterminacy with regard to fact, or with regard to epistemic ascertainability. Here I shall use terminology more appropriate to the latter.

[34] Ridge notes the possibility of different patterns by which a given level of acceptance might be realized, but confines his attention to cases in which the only question is which agents accept and which reject (but the number of each is held constant). He claims (implausibly, in my judgment) that in most of these cases the upshot is likely to remain the same (Ridge, 'How to Be,' p. 252).

[35] Of course this procedure would still be subject to the complaint that it is arbitrary to assume that all code-rejecters do the same thing, rather than allowing they would do diverse things, as would be more natural.

[36] Of course *accepting* and following a code will also generate positive psychological upshots, such as social approval and pride in oneself, but for simplicity of exposition I do not include such effects in this example.

[37] For helpful discussion and comments, I am very grateful to Douglas Blair, Pavel Davydov, Nancy Gamburd, Meghan Sullivan, and Evan Williams.