

REPLY TO ARNESON AND MCINTYRE

Brad Hooker
The University of Reading

Richard Arneson and Alison McIntyre have done me a great honor by reading my book *Ideal Code, Real World* so carefully.¹ In addition, they have done me a great kindness by reading it sympathetically. Nevertheless, they each find the book ultimately unconvincing, though in very different ways. But the cause of their dissatisfaction with the book is not mistaken interpretation. They have interpreted the book accurately, and they have advanced penetrating criticisms of it.

One group of their criticisms definitely draw blood. To treat the wound, my formulation of rule-consequentialism will have to be revised. A second group their criticisms seems to me fatal only if certain considerations are ignored. I will highlight the considerations that I think inoculate rule-consequentialism against these criticisms. In reaction to a third group of their criticisms, however, I have to accept that Arneson and McIntyre simply have quite different intuitions from mine, such that the prospects of agreement between the three of us are dim.

1. Arneson's Objection that Rule-consequentialism Is an Unstable Compromise

Arneson contends that rule-consequentialism is an unstable compromise between consequentialism and deontology. Arneson writes,

Either the deontological and other common-sense judgments that we are trying to accommodate should be accepted, in which case they should be reflected directly into the formulation of fundamental principle, or they should be resisted not accepted, explained away as an understandable mistake to which we are prone. But in that latter case the deontological intuitions pose no bar to acceptance of straight old-fashioned act consequentialism.

And again,

If one believes that consequentialism is wrong because it fails to accord with a morality of moral constraints and moral options, why not straightforwardly incorporate constraints and options directly into the formulation of fundamental moral principles?

Well, rule-consequentialism might be inadequate for many reasons, but I think not for this one. Why must a moral theory either give intuitively significant considerations a foundational role or none at all? True, the most familiar forms of act-consequentialism give no foundational role to agent-centered options or constraints. True, Rossian deontology does give them a foundational role. But we should be very unhappy about being pushed into choosing from a menu with only act-consequentialism and Rossian deontology on it.

The reason we should be very unhappy with such a restricted menu is that we seek a moral theory that (a) coheres with our considered moral convictions and (b) offers a unifying foundational principle that provides justification for our considered moral convictions (even, ideally, *impartial* justification for them). I take it that this is the heart of the search for reflective equilibrium in ethics.

Now we see the problem with the restricted menu that Arneson suggests. Act-consequentialism offers a unifying justificatory principle but not one whose implications most of us can believe. Rossian deontology has implications that most of us believe, but it offers no unifying justificatory principle. So act-consequentialism is routinely charged with being *highly counter-intuitive*, and Rossian deontology is accused of picturing morality as an *unconnected heap* of duties.

I think it is because each of the two options on that restricted menu seems unsatisfactory that so many moral philosophers have sought to come up with more options, such as Kantian ethics, contractualism, and virtue ethics. Hence I see the landscape in normative ethics as follows:

	Act- conseq	Rossian Deontology	Rule- conseq	Kantian ethics	Contract- ualism	Virtue ethics
Does the theory have <i>intuitively acceptable implications</i> ?	No	Yes	?	?	?	?
Does the theory offer a <i>unifying justificatory principle</i> ?	Yes	No	Yes	Yes	Yes	Yes
Does the theory's <i>unifying justificatory principle have intuitively acceptable implications</i> ?	No	Not applicable	?	?	?	?

One interesting feature of the theories mentioned in the above table is that, in their criteria of rightness, act-consequentialism and Rossian deontology are alike in having a one-level structure. Rule-consequentialism, contractualism, and virtue ethics clearly have a two-level structure: they each have a principle that selects rules or virtues, which then determine moral wrongness. In the relevant sense, Kantian ethics has a two-level structure as well. Indeed, each of the theories to the right of Rossian deontology in the above table seems to me to be trying to do precisely what Arneson complains about. Each of them tries to account for moral constraints and options without having them incorporated directly into the fundamental moral principle.

2. Arneson's Objection that Rule-consequentialism Doesn't Focus on the Right Reasons

A closely related objection to the one I have just considered is the objection that, even if rule-consequentialism gets lucky enough to come out with the right conclusions about what to do, it does so by the wrong route. In short, rule-consequentialism doesn't focus on the real reasons that things are right or wrong. Arneson writes,

Act consequentialism and common-sense deontology share a feature that is attractive, and that rule consequentialism lacks. If the question is raised, what morally should a particular person Smith do in a particular situation, the answer, according to both act consequentialism and common-sense deontology, is determined by qualities of the possible acts that the person could perform in this situation and perhaps by comparisons among the possible acts the person could perform. . . . According to both doctrines, what would happen if other agents did other acts on other occasions, and what right-making and wrong-making characteristics might attach to the possible acts that other agents might do on other occasions, are entirely irrelevant, and have no bearing whatsoever on the determination of what Smith morally ought to do here and now.

Arneson seems to me to overdraw the difference between common-sense morality and rule-consequentialism. Suppose you are in a circumstance where you could marginally improve your prospects for an enjoyable lunch by breaking your promise to meet boring me for lunch. What does common-sense morality say? It says you shouldn't break your promise for such a small gain in utility. And that is exactly what rule-consequentialism says as well. At least initially, common-sense morality and rule-consequentialism direct your attention to precisely the same features.

The real difference between them is that common-sense morality stops there, has nothing more to say, offers no further justification of the duty to keep one's promises. Rule-consequentialism, in contrast, has an explanation

of why there is a duty to keep one's promises. That is a huge attraction of rule-consequentialism. Again, rule-consequentialism aspires to provide a picture of morality according to which, although morality does contain various duties, they are *not* merely an unconnected heap, but instead have an underlying unifying rationale. And, again, this aspiration is an attraction that I think the best versions of Kantianism, contractualism, and virtue ethics share with rule-consequentialism.

But doesn't rule-consequentialism imply that what makes your standing me up at lunch wrong is not *really* that it constitutes breaking your promise, but *rather* that it is forbidden by the code whose internalization has the highest expected value? To be sure, rule-consequentialism purports to pick out the *ultimate* basis of right and wrong. But I don't think most people's intuition insists that reference to promise-breaking or lying or stealing or the other things picked out by Rossian deontology are necessarily moral bedrock.

3. Arneson's Objection about Having Different Rules for Different People

Arneson gives us the case to consider in which one person is very much smarter than everyone else. My rule-consequentialism implies that what this smart person ought to do is determined by the code whose internalization by 90% of everyone in each future generation has the greatest expected value. Moral rules that would require very complex calculations do not have high expected value, given that these difficult-to-apply rules would have to be learned and used by people who are not very smart. So the difficult-to-apply rules are rejected by rule-consequentialism. But doesn't this imply that the smart person must follow easy-to-apply rules even though *she* is smart enough to follow difficult-to-apply ones?

One could formulate rule-consequentialism so that it allows relativization to groups. Many rule-consequentialists have done so. The rule-consequentialist principle could make wrongness a function of the code whose internalization by *the agent's own group* has the highest expected value. If formulated this way, rule-consequentialism could allow one code for smart people and another one for cognitively challenged people. It could allow one code for rich and another for poor. Maybe it could allow one code for Americans, another for Canadians, another for Swedes, another for Germans. Indeed, why not one code for people from La Jolla, and another for people from Laguna Beach? Hell, why not one code for Bob, another for Carole, another for Ted, another for Alice? This runaway relativization flies in the face of the idea that the same rules should apply to everyone.

Much of the pressure towards relativization dissipates in the face of the following point. A code for internalization by everyone will distinguish between different situations in which people might find themselves. For

example, it could say that parents have responsibilities to the young that others don't have. It could say that those in a position to help the needy have a responsibility that those not in such a position don't have. It might even say that those who are very smart should work out the answers to certain complicated questions that others shouldn't feel obligated even to address. However, because of the extra costs involved in learning and remembering complications in a code, there is some limit on how many of these complications will be cost-effective to incorporate.

Let me distinguish between two questions. One is: how many such complications are contained by the code with the highest expected value given that this code is to be internalised by such-and-such group, e.g. by the group constituted by 90% of each new generation? That is a consequentialist question downstream from a decision about whether to relativize codes to groups. The upstream question is, how many different groups should humanity be divided into so that we can relativize codes to groups? I do not answer that upstream question in a consequentialist way. The answer to it is built into my formulation of the rule-consequentialist foundational principle.

However, another of Arneson's objections shows that my theory needs revision on this matter. Again, my theory makes moral wrongness a function of the code of rules whose internalization by 90% of each new generation has the highest expected value. Arneson asks us to imagine that a thousand years from now there are technological breakthroughs that enormously reduce the costs of thereafter inculcating in new generations much more complex and demanding codes. Arneson observes, "Rule consequentialism makes the determination of what is right here and now hostage to contingencies concerning what rules would produce good consequences if internalized by future people in whatever circumstances those future people happen to face. Intuitively those contingencies do not seem to be determiners of right and wrong."

He is right. My theory as formulated has the result he says it has, and that result is implausible. And I see no escape from his objection that does not require revision of my formulation of rule-consequentialism—and revision so as to require a degree of relativization.

When in the book I first give a full statement of rule-consequentialism (*ICRW*, p. 32) I included a footnote saying that the cost-benefit analysis of various possible codes is to be run on the assumption that new generations are not changed by genetic engineering. Arneson's objection reveals that I should not have drawn the horizon at the advent of genetic engineering. The horizon should instead have been drawn at the advent of any new development that significantly reduces the costs of internalising more complex or demanding codes. Thus my revised principle would be:

Moral wrongness is determined by the code of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation (not

including generations after any new development that significantly reduces the costs of internalising more complex or demanding codes) has maximum expected value in terms of well-being with some priority for the worst off.

4. Arneson's Objection about Publicity

Arneson thinks I am wrong to reject the idea that true morality could be esoteric, and wrong to think rule-consequentialism must incorporate publicity of the ideal code.

My book argued that esoteric morality is rejected by rule-consequentialism (*ICRW*, pp. 85–6). Here was my argument:

Statement of rule-consequentialism	Rule-consequentialism holds that moral wrongness is determined by the code whose internalization by 90% of everyone has maximum expected value.
Sociological claim	One consequence of 90% of everyone's internalizing a given code would be public knowledge of the widespread internalization of that code.
Conclusion	Any code that rule-consequentialism endorses will be suitable for public recognition.

That argument could be unpacked as:

Sociological claim	One consequence of 90% of everyone's internalizing a given code would eventually be public knowledge of the widespread internalization of that code.
Conceptual claim about expected value	If one consequence of 90% of everyone's internalizing a given code would eventually be public knowledge of the widespread internalization of that code, then the expected value of 90% of everyone's internalising a given code includes the expected value of eventual public knowledge of this.
So	The expected value of 90% of everyone's internalising a given code <i>includes</i> the expected value of eventual public knowledge of this.
Statement of rule-consequentialism	Rule-consequentialism holds that moral wrongness is determined by the code whose internalization by 90% of everyone has maximum expected value.
Conclusion	Rule-consequentialism maintains that moral wrongness is determined by the code whose internalization by 90% of everyone has maximum expected value where this <i>includes</i> the expected value of eventual public awareness of this code's widespread internalization.

Obviously, the argument above has three premises: (1) a sociological claim, (2) a conceptual claim about expected value, and (3) the statement of rule-consequentialism. Arneson is right that the rule-consequentialist argument for publicity draws on more than merely the statement of rule-consequentialism itself. But I can't see how the conceptual claim about expected value can be denied. And the sociological claim just does seem true. So I'm happy with the conclusion that rule-consequentialism embraces publicity for its rules.

But what if the sociological claim is not true? Arneson is right that I want to make the connection between rule-consequentialism and publicity so tight that it doesn't really rely on such sociological claims. This is because, like Kantians and contractualists, I think of morality as a system aimed at impartial, indeed interpersonal, justification. To make this more prominent, perhaps my formulation of rule-consequentialism should have referred not only to widespread internalization of rules but also to public awareness of their widespread internalization.

5. Arneson's Objection about Non-compliance

Rule-consequentialism had better be formulated so that it not only says what acts are wrong in full compliance situations but also says what acts are wrong in partial compliance situations. If rule-consequentialism says that, even in partial compliance situations, we should follow the rules that would be best in full compliance situations, then rule-consequentialism is absurd. So rule-consequentialism had better specify some other set of rules to follow in partial compliance situations.

That is a massively important point. My book tries to be sensitive to it—indeed, the very title of the book is meant to suggest sensitivity to it. More substantively, my rule-consequentialist principle is formulated so as to evaluate possible rules in terms of the expected value of their internalization by 90% of the population in new generations. So rule-consequentialism's principle picks a level of non-acceptance by the whole population as part of the specification of the conditions under which alternative possible rules are assessed. This recognizes the need for rules dealing with partial compliance. (It even recognizes the need for rules dealing with complete amoralists.)

Aware of those points, Arneson objects that "sophisticated rule consequentialism with its disaster avoidance component tells us to obey the ideal code of rules in scenarios in which common sense morality would rebel from this conclusion." Arneson writes,

Consider situations in which the ideal code of rules, or at least the portion of it that is in question on this occasion, is not in fact accepted by most people and not followed by most people. Consider a rule that would produce ideal

consequences if everybody or nearly everybody conformed their behavior to it, but would produce no good consequences otherwise. Here is a simple example: In war, soldiers fighting for a just cause ought to stand by their post when attacked, unless outnumbered by attacking enemy so that even stout defense would be futile. Suppose this rule, followed by nearly everybody, would produce ideal results. But the rule in fact is not internalized by the military forces fighting for a just cause in a particular war. The enemy have attacked and most of your fellow troops have run away. You can stand and fight, in conformity with the ideal rule, or you can run and live to fight another day. The consequences of conformity to the rule would not be disastrous, but would be decidedly negative. You will die and gain very little if anything for your side... Common sense morality, which holds that the obligation to obey hypothetically useful rules is sensitive to the actual degree to which others are complying here and now, surely says one should run and live to fight another day. Act consequentialism to its credit says the same. Rule consequentialism, even sophisticated rule consequentialism with the disaster avoidance proviso added, would have to hold that one ought to stand and fight and die. So much the worse for sophisticated rule consequentialism.

Just before Arneson advances this objection, he concedes that rule-consequentialism is in agreement with common-sense morality in holding (against act-consequentialism) that

the moral obligation to tell the truth, keep one's promises, and in general to conform to significant moral rules continues to hold and should constrain the conduct of the morally conscientious agent even when lying or breaking one's promise or the like would bring about somewhat more good than standing fast by the moral rules.

The objection Arneson is now advancing is not *blankly* that rule-consequentialism requires the agent to stick to the rules, but rather that rule-consequentialism requires the agent to stick to the rules *even when others are not*, and so no (or virtually no) good would actually result.

At the end of his paper, Arneson modestly suggests that his criticisms of my theory are "variants of standard criticisms". I think some of his criticisms are in fact wholly new; however, the one now under consideration really is a variant of a standard criticism. It is a variant of one of David Lyons's criticisms of rule-consequentialism. Consider a rule of which both of the following are true. Your following this rule will typically be burdensome for you. Everyone's internalising this rule will maximize expected value. Lyons's criticism was that rule-consequentialism requires you to follow such a rule even when this rule is being ignored by others.²

My book admits that this criticism of rule-consequentialism is extremely important—in other words, that rule-consequentialism had better have a good answer to it! The answer to this criticism that the book offers is,

Rule-consequentialism must not require agents to make sacrifices for others who are able to follow the same rule but won't... To discourage people from becoming or remaining free-riders on the kindness or restraint of others, the ideal code does not require kindness or self-denying restraint towards people not disposed to reciprocate. (Exceptions will naturally be made for obligations towards those of diminished responsibility, such as young children.) (*ICRW*, p. 125)

The implication of this is that rule-consequentialism underwrites a qualification to more or less all the rules it endorses, a qualification specifying that one is not required to restrain oneself or make sacrifices for the benefit of free-riders.

In order to consider Arneson's example in the light of that qualification to rules, we need to ask whether the main beneficiaries of your standing by your post would be your fellow soldiers, who are in fact not standing by their posts? They are likely to be the *temporally* first beneficiaries. Your standing by your post would hardly change the course of the war, but it is likely to hold up the enemy's advance at least a little. And holding up the enemy's advance increases the likelihood that your fellow soldiers will escape to increased safety, or at least live that much longer. So the principle about reciprocity I offered does seem applicable here. You aren't required to stand by your post to protect others who have already indicated by their actions that they won't do the same for you.

But now suppose we tweak Arneson's example so as to make irrelevant the rule (or more accurately a qualification to all rules) intended to discourage free-riding. Suppose your fellow soldiers would not in fact benefit in any way from your standing by your post. Actually, I think this is what Arneson had in mind. As he frames the issue, "You can stand and fight, in conformity with the ideal rule, or you can run and live to fight another day. The consequences of conformity to the rule would not be disastrous, but would be decidedly negative. You will die and gain very little if anything for your side."

I share Arneson's intuition that, if standing by your post will result in your immediate death and virtually nothing for your (just) side, then morality allows you to join your fellow soldiers in fleeing. What I don't see, however, is why your standing by your post under these conditions wouldn't count as disastrous. Given that you have much to live for, your being immediately killed would be disastrous *for you*. Isn't the disaster to you enough for the rule about disaster prevention to kick in?

As far as I know, everyone who has commented on my book has accepted that rule-consequentialism can and should prescribe a disaster prevention rule. Admittedly, what counts as a disaster is vague. It is also variable (for example, what counts as a disaster the prevention of which would justify breaking a promise is far less than a disaster the prevention of

which would justify imposing serious physical harm on someone). I also admit I didn't say enough in the book about disaster prevention when the disaster is to oneself. What I should have done is to say that disaster prevention is an *obligation* when the disaster is to *others*, and a *permission* when the disaster is to *oneself*.

On the other hand, I did note that the disaster prevention rule can override deontological constraints where the disaster would be to oneself. I gave the example, "my duty to tell the truth . . . may stop short of requiring me to reply honestly to the Inquisition's question about whether I am an atheist." (*ICRW*, p. 165, fn. 10). Perhaps, equally, you are not required to stand by your post when this would do your side no good but would be disastrous for you.

There are, of course, further questions about the extent to which soldiers, fire-fighters, rescue services, police, and the like can be morally required to undertake bigger risks for the sake of saving others than ordinary people are. But rather than try to wade into those questions, I will move on to Arneson's other criticisms.

6. Arneson's Objection about Giving Foundational Importance to Fairness

Arneson's next criticism of my theory has to do with what I say explicitly about fairness in the context of compromising with conventionally accepted rules. I wrote, if rule-consequentialism is to be plausible, "we need a rule requiring us to comply with conventional sub-optimal rules when our not doing so would cause serious unfairness." (*ICRW*, p. 122). Arneson comments that this addition transforms my theory into "a hybrid or compromise", not really pure rule-consequentialism. He writes, "If fairness is a nonconsequentialist value that matters morally, I don't see how its writ can be restricted to conditions of general nonacceptance. We are on the road to affirming some version of pluralist intuitionism, not any version of rule consequentialism." Here, again, Arneson definitely draws blood.

I defended rule-consequentialism by adding to it a principle "Behave fairly" for dealing with situations where others do not accept the ideal code. In effect, this defensive move puts into action a principle about behaving fairly in all the contexts where others do not in fact accept the ideal code. And such contexts may well be the norm rather than the exception. So Arneson is correct that my attempted defence of rule-consequentialism actually only sidelined rule-consequentialism.

But the problem for my theory might be even deeper than that. For suppose that on Monday the code whose internalization has the highest expected value is one whose principle about contracts is a rule let's call the Monday Contract Rule. And suppose on Monday you and I signed a partnership contract that was binding according to the Monday Contract

Rule. Suppose also that for some reason on Tuesday the code whose internalization has the highest expected value is one whose principle about contracts is different, such that our contract would not be binding according to the Tuesday Contract Rule. (For example, the Monday Contract Rule does not require us each to be advised by a lawyer in order for the contract to be binding; the Tuesday contract rule does require each of us to have legal advice.) Tuesday arrives and you wonder whether you are morally required to comply with the agreement you made on Monday.

This problem for rule-consequentialism repeats. We sign a contract on Tuesday that is binding according to the Tuesday Contract Rule. But suppose that on Wednesday for some reason the code whose internalization has the highest expected value is one whose rule about contracts is a rule that conflicts with Tuesday's Contract Rule. (For example, Wednesday's rule stipulates a minimum period for reflection between circulation of a proposed contract and its acceptance.) Likewise, suppose that on each subsequent day a new rule about contracts has higher expected value thereafter than the previous day's rule has.

This problem for my theory cannot be handled merely by adding a qualification to all rules to discourage free-riding. The cases I'm worried about here are not ones where anyone is setting out to free-ride. The problem for my theory is that the code whose internalization has the highest expected value might change with each new day (indeed, minute!). Since rule-consequentialism claims that moral requirements are determined by the code with the highest expected value, rule-consequentialism seems sharply at odds with the intuition that moral requirements cannot be so wholly future-oriented and fluctuating. What we have agreed to in the past can be morally decisive, even if we now or later see that a new code would be better than the one we accepted in the past.

My book's infuriatingly vague answer to this problem is: "Be fair to yourself and others". This answer is so vague because 'fair' is used so broadly. For example, in everyday moral talk, 'fair' is often used to mean "in conformity with all relevant moral distinctions". In other words, on this usage, "Be fair" means "Do what is morally required, all-things-considered". But, in this sense, "Be fair" doesn't at all tell us *what* morality requires; it merely tells us *to do* whatever morality requires.

There are a number of more restricted (and useful) ways of using the term 'fair'. I myself tentatively think that, on the best narrow usage of the term, acting fairly has mainly to do with honoring agreements, respecting desert, and giving priority to the worst off. Furthermore, protecting the force of agreements is the main element in the problem that I was worried about in the book when I invoked fairness in a way that Arneson pointed out is ad hoc.

Return to the case where the code with the highest expected value on Tuesday differs relevantly from the code with the highest expected value on

Monday. Common-sense morality says that on Tuesday we should abide by the contracts we signed on Monday, given that the contract was binding according to the code with the highest expected value on Monday. Can rule-consequentialism say the same thing, without resort to ad hoc devices?

Rule-consequentialism had better have some way of dealing with the fact that the code with the highest expected value on Monday can differ from the code with the highest expected value on Tuesday, which code can differ from the code with the highest expected value on Wednesday, etc. If the theory doesn't have some way of dealing with code variation over time, then it self-destructs. One thing rule-consequentialism is especially keen on is assured expectations. If you know on Monday that the contract you are asked to sign with me that day may not be valid on Tuesday or Wednesday or whenever because Monday's code will be replaced on Tuesday or Wednesday or whenever, then you won't have much assurance that Monday's contract will be binding beyond Monday. That is equally true of any contract you are asked to sign on Tuesday, and indeed of any contract to be signed on *any* day.

To generate the needed assurance, a code would have to contain a rule putting pressure on people to abide by the agreements they made if those agreements were valid according to the rules that were mutually accepted by the parties at the times they made the agreements. Of course, there will have to be some limits on the kinds of agreements held to be binding (e.g. no slavery contracts). But, apart from some few such limitations, the rule about agreements would be that they are binding if they were valid according to the rules accepted at the time the agreements were made. So an old agreement can be binding even if a new code that is now seen to have the highest expected value lays down different rules for making new agreements. Such an approach answers the specific objection I had in mind when in the book I wrongly invoked "Be fair".

Other objections to rule-consequentialism may emerge from the theoretical possibility that everyday there is a new code with the highest expected value. Considering those objections, however, must be the task for a different occasion.

7. McIntyre's Objection about Sadistic Pleasures

While Arneson accuses rule-consequentialism of focusing on moral irrelevances, McIntyre accuses it of being willing to count in favour of some things that we might think actually count against it. She writes, "some contributions to aggregate well-being simply do not count from a moral point of view: the entertainment value of public executions, the pleasures of scapegoating, mockery, and public vilification of wrongdoers, and the pleasures of self-congratulation."

McIntyre is right to raise the issue of evil pleasure. But I think she takes the category of evil pleasures more broadly than I do. I want to disassociate sadistic pleasure from the pleasure of self-congratulation, something I am not so puritanical as to think always bad. Sadistic pleasure, as I understand it, is pleasure taken in the suffering of others, whether or not that pleasure be sexual in nature.

In my book, I unwittingly ducked sadistic pleasure. Even now, I'm not sure what to say about it. As far as I can see, there are available to me only three possible responses to McIntyre's objection about sadistic pleasure.

The first of the three possible responses is for me to admit complete defeat. Suppose sadistic pleasures *do* increase the well-being of the person who experiences them. Suppose these additions to well-being *are* counted in the rule-consequentialist cost-benefit analysis of alternative possible codes. Suppose the cost-benefit analysis *could* come out favouring say ritual torture or capital punishment precisely because of these pleasures. I concede to McIntyre that such a result would render rule-consequentialism very unattractive.

The second possible response to the problem of sadistic pleasures starts by again accepting that sadistic pleasures *do* increase the well-being of the person who experiences them. This second possible response continues by accepting that these pleasures *are* to be counted in the cost-benefit analysis of alternative possible codes. But this second possible response *denies* that the cost-benefit analysis *would* come out favouring say ritual torture or capital punishment. Stimulating the appetite to take pleasure in the suffering of others is horribly corrosive, in that it works against universal harmony and global benevolence. So any code acceptable to rule-consequentialism will suppress, or at least sublimate, sadism rather than feed it.

That is probably what I was assuming when I wrote *Ideal Code, Real World*. It is a comfortable assumption for a defender of rule-consequentialism to make. Therefore it is an assumption about which I should be sceptical.

The third possible response available to me is like the first two in accepting that codes should be assessed by effects on well-being. But this third response denies that sadistic pleasure can add to well-being. We might say the resulting position assesses alternative possible codes in terms of their effects on "moralized well-being". (Incidentally, this position is roughly the one Korsgaard ascribed to Bernard Williams.³ He confirmed in conversation that it was his view, though he wanted to "avoid letting it sound too consequentialist".)

Let me acknowledge the advantage and disadvantage of this third line of response. The advantage is that it *guarantees* that the production of sadistic pleasures never counts in favour of a code. The disadvantage is that it involves adding a stipulation right at the foundation of rule-consequentialism, a stipulation that a maximally ambitious form of

rule-consequentialism would not make. The stipulation that sadistic pleasures are not to count in the rule-consequentialist assessment of alternative possible codes significantly weakens the explanatory power of rule-consequentialism. If sadistic pleasure is ignored in rule-consequentialist assessment of possible codes because sadistic pleasure is morally wrong, then this is a kind of moral wrongness that rule-consequentialist assessment must presuppose rather than try to explain.

I am unsure which of the three possible lines of response to take. I'm far from entirely happy with any of them. Like most other consequentialists, however, I think that on balance the second alternative least bad.

8. McIntyre's Objection that Predictability Is Rare and Assessment Piecemeal

Further discomfort is inflicted on me by McIntyre's complaint that my appeals to the ideal code turn out to be a mere distraction. I accept that predicting the consequences of social change with reasonable confidence is extremely difficult. And, typically, the greater the number of simultaneous changes of which one is trying to predict the consequences, the more difficult the prediction is. In particular, because of the difficulty of predicting the consequences of across-the-board replacement of a moral code, I retreat to the view that the currently accepted moral code should be revised if and only if such revisions have greater expected value than sticking with the status quo. But this, as McIntyre writes,

involves abandoning the aim of explaining, justifying or revising any of our actual moral convictions by reference to the content of an ideal moral code. Piecemeal improvements in the "efficiency" of a current code cannot be cast as approximations to an ideal one.

McIntyre is right. To see how troubling her objection is, consider an analogy. Suppose I want to wear whatever outfit is optimal for me. The procedure analogous to the one rule-consequentialism seems to be pushed back to is this. I should think of my whole outfit apart from my tie as fixed and then evaluate which tie to wear. Suppose that, given what I have on already, an orange tie would be best. Then I could think of all my outfit except my shirt as fixed and evaluate which shirt to wear. A blue shirt is best given what else I have on. Then I take my whole outfit except my jacket as fixed. I choose my navy jacket. Then I take my whole outfit except my trousers as fixed and evaluate which trousers to wear. I choose grey trousers. So now I have on an orange tie, blue shirt, navy jacket, and grey trousers. This outfit might not be too terrible. But my dark blue suit, striped shirt, and maroon tie would have been a far better package. Very

depressingly, notice that no element of the package arrived at via piecemeal assessment appears in the ideal package.

I admit we don't know what the ideal code is. I admit, as she quotes me, that we "would be silly ever to think we have found a completely unimprovable code". Now I have to add the admission that we can't even be sure that the code with the highest expected value, by our lights, contains *at least some* rules that would appear in the ideal code. Perhaps my book's title *Ideal Code, Real World* was too idealistic. Perhaps a better title would have been *Incrementalism: Making Do With Indeterminably Nonideal Rules*.

In the face of objections about the unpredictability of more radical reform, I have fallen back to a piecemeal, incrementalist approach. This incrementalist approach doesn't have the idealistic chime of appeal to the ideal code. That seems to me a real loss. Nevertheless, modest as the incrementalist approach is, it may be the most justified approach to moral reform.

9. McIntyre's Objection about Priority for the Worst off

My formulation of rule-consequentialism claims that wrongness is determined by the rules whose internalization has "maximum expected value in terms of well-being (with some priority for the worst off)". (*ICRW*, p. 32) McIntyre comments, "the consequentialist warrant for including such priority is not made clear".

She is right. But, at the level in which I want to insert priority for the worst off, there could be no consequentialist warrant for it. For the level in which I want to insert it is the foundational level, the level of first principle. There is no deeper level of a normative moral theory than its first principle. (There is a deeper *metaethical* level containing principles about how to evaluate normative theories. In section 1 above, I alluded to some of the pivotal metaethical principles.)

Consequentialist evaluation of rules evaluates them by their consequences. Which consequences matter in such evaluation? Consequentialism itself does not answer that question. Rather, consequentialism needs that question already answered. Suppose the answer is *consequences for well-being*. But how is well-being to be calculated? The traditional answer, of course, is *impartially*—i.e., benefits or harms to any one individual are counted the same as the same size benefits or harms to any other individual.

However, because of the counter-intuitive implications of that utilitarian approach, I tentatively favour calculating well-being by giving extra importance to the plight of the worst off. So my reason for such priority is not consequentialist. Rather, it is that doing so makes rule-consequentialism more likely to prove intuitively acceptable.

10. McIntyre's Objection About Directing Resources to the Disabled

A related objection of McIntyre's concerns diverting resources from the able-bodied to the disabled. She writes,

[T]he severely disabled may be assumed to be inefficient converters of resources into utility or well-being in so far as they require more resources in order to have the kinds of educational, employment, and independent living opportunities that are comparable to those enjoyed by nondisabled people. On this assumption, it follows that diverting more resources toward the goal of providing such opportunities for disabled people would inevitably result in a drop in aggregate well-being.

Despite the drop in aggregate well-being, where this is calculated in a strictly impartial way, not in a prioritarian way, McIntyre thinks that such a diversion of resources would be a moral improvement. However, if the disabled are the worst off, then, even if diverting more resources to them reduces aggregate well-being calculated impartially, diverting more resources to them may be compatible with a form of rule-consequentialism that calculates expected value by giving priority to the well-being of the worst off.

And what if the disabled are not the worst off? Then prioritarian rule-consequentialism might prefer a code that does less for the disabled and more for those who are worst off. This seems to me intuitively right.

A different objection is suggested in the final paragraph of McIntyre's paper. That paragraph poses a choice between two possible social policies.

One of these social policies does *not* divert greater resources to the disabled. Thus, resources are employed by people who get larger benefits from them. So, with this social policy, there is greater impartially-calculated aggregate well-being in the current generation. In addition, because resources are not diverted to the disabled, there is an increase in people's electing to undergo genetic screening and selective abortion for physical disabilities. The result is a decrease in the percentage of the people in future generations who are inefficient converters of resources into well-being.

The alternative social policy *does* divert greater resources to the disabled. The result is a lower level of impartially calculated aggregate well-being in the current generation, because more resources go to people who are inefficient converters of resources to well-being. Another result will be a lower level of well-being in future generations, because the percentage of people in future generations who have disabilities will be higher.

McIntyre clearly thinks the social policy of diverting greater resources to the disabled is morally superior. In so far as the disabled already exist and are the worst off, then prioritarian rule-consequentialism can agree. The intriguing thing about McIntyre's thought experiment, however, is that the

choice between the two social policies is partly a choice of which set of people will come into existence. She rightly thinks that, in so far as we are focusing on future generations, rule-consequentialism must favour bringing about people who are better at converting resources into well-being. Again, this seems intuitively right to me.

11. Diverging Intuitions

At least with respect to that question of social design, McIntyre has intuitions that clearly part company with the kind of rule-consequentialism set out in my book. Her intuition about the above case gives her good reason to reject any form rule-consequentialism that evaluates codes in terms of well-being, even prioritarian rule-consequentialism.

Arneson, too, has intuitions that lead away from rule-consequentialism. He has the intuition that what is wrong for an *individual* to do can't be a function of what code it would be best for some *group* to internalize. Yet one of rule-consequentialism's central ideas is that moral rules have to be designed for acceptance by the group. In addition, Arneson comments, "I do not myself believe that an adequate morality does include constraints and options." Yet my argument for rule-consequentialism partly depends on the theory's ability to underwrite agent-relative constraints and options.

Put crudely, Arneson and McIntyre have some important intuitions that march away from rule-consequentialism, albeit (I suspect) in opposite directions, Arneson's in an act-consequentialist direction, McIntyre's in a more deontological or perhaps virtue-based direction. This gives me all the more reason to be grateful to them for reading my book so carefully and sympathetically.⁴

Notes

1. *Ideal Code, Real World: A Rule-consequentialist Theory of Morality* (Oxford: Oxford University Press, 2000). Henceforth, my references to the book will refer to *ICRW* and be placed in the text.
2. David Lyons, *Forms and Limits of Utilitarianism* (Oxford: Oxford University Press, 1965), pp. 128–32, 137–42.
3. Christine Korsgaard, *The Sources of Normativity* (Cambridge, England: Cambridge University Press, 1996), pp. 74–5. See Bernard Williams, *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press, 1985), pp. 45ff., 147–54).
4. This paper has been presented at the 2004 Pacific Division Meeting of the American Philosophical Association, and at an Edinburgh conference on

utilitarianism organized by Elinor Mason and Mike Ridge, and at the University of Reading. For helpful comments, I am grateful to Jeppe Andersen, Elizabeth Ashford, Robert Audi, Sarah Buss, Thomas Carson, Tim Chappell, John Cottingham, Jonathan Dancy, Alice Drewery, Julia Driver, Leonard Kahn, Stephanie Lewis, David McCarthy, David Oderberg, Charlie Pelling, Mike Ridge, Geoffrey Sayre-McCord, John Skorupski, Philip Stratton-Lake, Jussi Suikkanen, Mark Young, Andrew Williams, and of course Richard Arneson and Alison McIntyre.