Jacob Ross

Consequentialism and Actual Rules

0. Introduction

After considering some problems facing exiting versions of consequentialism, I present the schematic outline of an alternative version of consequentialism, which I call *Actual Rule Consequentialism* (ARC). I then consider three ways in which one might attempt to ground ARC by deriving it from some more fundamental principle. The first attempt is to derive ARC from a principle of fairness, the second is to derive ARC from a principle of agential involvement, and the third is to derive ARC from a transmission principle connecting the reasons of groups to the reasons of group-members.

1. Problems with Act Consequentialism and Rule Consequentialism

While Act- and Rule- Consequentialism each face a number of objections, here I will focus on one particularly problematic objection facing each theory. Let's first consider

> **Act Consequentialism**: An action is permissible just in case the (expected) value of its consequences is at least as great as the (expected) value of the consequences of any alternative action available to the agent.

One objection to Act Consequentialism proceeds thus:

> The correct moral theory must be a moral theory we should want everyone to accept and try to follow. But we shouldn't want Act Consequentialism to be universally accepted, because such universal acceptance would have disastrous consequences. It would have such consequences because it would undermine a number of very valuable practices. Thus, for example, the practices of truth telling and promise keeping are very valuable. For our ability to rely on these practices greatly facilitates the communication of information and coordination of activities. Similarly, the practice of refraining from killing, harming and stealing is valuable, since it creates a sense of security. And the practice of prioritizing the interests of the near and dear is valuable, since it promotes close personal relations. But if everyone accepted Act Consequentialism, then people would lie, break promises, kill, harm, steal, etc., whenever doing so maximizes expected value, and no one would prioritize the near and dear. And so the universal acceptance of Act Consequentialism would undermine these valuable practices.

Call this the *Valuable Practice Objection*.

Now consider Rule Consequentialism. There are many versions of this view, but many of them can be represented by the following schema.

> **Rule Consequentialism**: An action is permissible just in case it is permitted by some rule R such that R's being X would have as good or better (expected) consequences than any alternative rule's being X.

Different values of X yield different versions of rule consequentialism. Thus if we substitute *universally complied with* for X, we get Universal Compliance Rule Consequentialism. And the same applies, mutatis mutandis, if we substitute *universally accepted*, or *universally followed* for X. Similarly, if we substitute *complied with by the majority* for X, we get Majority Compliance Rule Consequentialism. And similarly, mutatis mutandis, if we if we substitute *accepted by the majority* for X, or *followed by the majority* for X. The following case raises a problem for any such version of Rule Consequentialism.

> *The Machine of Doom*: There is a machine which ascertains two facts: first, which rules are X, and second, whether agent s follows rule R. If R is X, then the machine brings it about that everyone gets eternal bliss. If R isn't X and s follows R, then everyone gets eternal torment. And if R isn't X and s does not follow R, then everyone lives the usual kind of life without either eternal bliss or eternal torment. As a matter of fact, regardless of whether s follows R, R won't be X. And s knows all these facts.

Rule consequentialism implies that, in this scenario, it is permissible for s to follow R, and impermissible for s to follow any alternative rule. This implication, it seems to me, is highly implausible, regardless of how we fill in the details of the scenario. As an illustration, suppose rule R states "whenever there's a full moon, stand on your head and spit wooden nickels." Let X be the property of being accepted by the majority of the population, and let s be Fred. Hence, we must suppose that, regardless of whether Fred follows the full moon rule, this rule won't be accepted by the wide majority of the population. Clearly, in *The Machine of Doom*, this rule is such that its being accepted by the majority would have better consequences than any alternative rule's being accepted by the majority. And since this rule permits standing on one's head and spitting wooden nickels when there's a full moon, Rule Consequentialism implies that it would be permissible for Fred to do so. But since Fred knows that his doing so would result in everyone's undergoing eternal torment, it seems clear that it would not be permissible for Fred to do so.

This example shows that there are possible worlds in which Rule Consequentialism gets the wrong results. We can also show that, at least for popular ways of specifying the value of X, Rule Consequentialism gets the wrong results in the *actual* world.

I can give a better example. Let R be a rule that people don't follow, but that it would be great if everyone followed. Let R* be the conjunction of R and a rule requiring answering yes to

the question "does everyone follow R?" R* will thus be optimific. Now suppose someone asks you "does everyone follow R?" R* will say to answer yes. So RC says it would be permissible to answer yes.

Consider the following interrogative sentence-schema:

IST: "Does the property of being X belong to a rule which requires answering 'yes" to this very interrogative sentence?"

And consider the following rule-schema:

RS: If asked question IST, answer 'yes.'

For the standard values of X, in the actual world, the answer to IST is "no." Suppose, for example, that X is the property of being accepted by the majority. IST is therefore "does the property of being accepted by the majority belong to a rule which requires answering 'yes' to this very interrogative sentence?" Clearly, it is not the case that the majority accepts a rule requiring answering 'yes' to the sentence above. And so, in the actual world, if one were asked IST, one should answer 'no.' But it follows from the version of Rule Consequentialism under consideration that, in the actual world, it would be permissible to answer IST affirmatively. For if RS were accepted by the majority, then the answer to IST would be 'yes,' and so there would be no problem with the large majority accepting RS. Thus, it seems that RS is a rule whose acceptance by the majority would have as good or better consequences as the acceptance of any alternative rule by the majority. And since RS permits answering "yes" to IST, it follows from Rule Consequentialism that it is permissible, in the actual world, to answer "yes" to IST.

The problem I have been discussing is a familiar one. There are plenty of rules such that, if everyone, or the most people, accepted or followed them, things would be hunky-dory, but which are also such that, in a world where it is not the case that everyone, or most people, accepts or follows them, then a given agent should not follow them, since her doing so would have bad consequences. Following Parfit, we can call this the *Ideal World Objection*.

It's worth noting that the ideal world objection is a problem for many other moral theories apart from Rule Consequentialism. Famously, it's a problem for the Categorical Imperative. But it's also a problem for any version of contractualism of the following form:

> **Contractualism**: It is permissible for follow a rule R just in case (if under ideal conditions one were choosing which rules will be X, one could rationally or reasonably choose that R be X).

For, in the world described in Machine of Doom, under ideal conditions one could rationally and reasonably choose that property X belong to the rule requiring standing one one's head and spitting wooden nickels. And in the actual world, one could, under ideal conditions, rationally and reasonably choose that property X belong to RS. And so Contractualism implies that it would be permissible to stand on one's head and spit wooden nickels in Machine of Doom, and it likewise implies that, in the actual world, it would be permissible to answer IST affirmatively.

2. Actual Rule Consequentialism to the Rescue

In this section I will propose an alternative form of consequentialism. This alternative will be a principle about the strength of moral reasons, rather than a principle about moral permissibility. However, since there is clearly a close connection between moral permissibility and the strength of moral reasons, this principle will bear upon questions of moral permissibility, and given the right background information, it may enable us to answer such questions. Once again, I will present this principle schematically, as follows.

> **Actual Rule Consequentialism (ARC)**: There is reason of strength Z to follow rule R whenever (i) R is in fact X, and (ii) the fact that R is X has sufficiently good consequences. There is no similar reason to follow R when either (i) or (ii) is not satisfied.

(Note: I'd be very interested to know who else endorses something like ARC. I know a principle of this kind is defended by Conrad Johnson in *Moral Legislation: A Legal-Political Model for Indirect Consequentialist Reasoning*).

It seems plausible that a principle of this form could avoid both the Valuable Practice Objection and the Ideal World Objection. Consider first the Valuable Practice Objection. Recall that Act Consequentialism is vulnerable to this objection because there appear to be a number of valuable practices (truth telling, promise keeping, prioritizing the near and dear, etc.) that would be undermined if everyone accepted Act Consequentialism. By contrast, everyone could accept Actual Rule Consequentialism without any of these practices being undermined. Consider, for concreteness, the version of ARC on which property X is the property of being accepted by the majority. This version of ARC says that we have reason to follow a rule whenever this rule is accepted by the majority and this fact has sufficiently good consequences. But in the case of the valuable practices undermined by Act Consequentialism, what underlies these practices are rules that are accepted by the majority, and where this fact has very good consequences (e.g., the rule requiring promise-keeping). And ARC implies that we have reason to follow these very rules. And so ARC supports, not undermines, these valuable practices.

Next Consider the Ideal World objection. Recall that Rule Consequentialism, the Categorical Imperative, and versions of contractualism all share this problem, because they all imply that it is permissible to follow a rule R whenever R's being X would have optimal consequences, or whenever R's being X could be rationally or reasonably willed, *even when* R is

not in fact X, and *even when*, in a world where R is not X, following R would be disastrous. Actual Rule Consequentialism avoids this problem. For it is only when a rule has property X that ARC implies that we have reason to follow this rule. According to ARC, if a rule R lacks property X, then the mere fact that wonderful consequences *would* result if R *were* X gives us no reason to follow R. Thus, ARC does not imply that in *The Machine of Doom*, one has reason to follow the rule requiring standing on one's head and spitting wooden nickels. And, similarly, ARC does not imply that, in the actual world, we have reason to follow the rule that prescribes answering "yes" to the self-referential question IST.

So far I have presented only a schematic outline of ARC. In order to arrive at a formulation of this principle that is practically adequate, we would need to answer three questions. First, how are we to specify the value of X, as it figures in ARC? That is, does the most plausible version of ARC concern rules that are complied with, or accepted, or followed, or what? And how widely must rules be complied with or accepted or followed (or what have you) in order for these rules to fall within the scope of ARC? Second, what counts as *sufficiently good consequences*? In order for R's being X to have sufficiently good consequences, must R be optimific, in the sense that there is no alternative rule R* such that R*'s being X would have better consequences than R's being X? Or could sub-optimific rules count as sufficiently good, and hence fall within the scope of ARC? And third, how are we to specify the value of Z? That is, on the most plausible version of ARC, how strong is the reason provided by a rule R when R is X and when R's being X has sufficiently good consequences? And does the strength of this reason depend on the degree to which R is X, or on how good the consequences are of R's being X? An adequate formulation of ARC would have to answer all these questions.

In addition to facing these questions about the proper formulation of ARC, the defender of this principle also faces an explanatory question. If ARC is true, then why is it true? There may be some practical principles that are so self-evident that they could plausibly be regarded as explanatorily fundamental, neither requiring nor allowing for any deeper explanation. But ARC does not appear to be such a principle. If it is indeed true, then we should hope to be able to say something about why it is true.

I believe that these questions of formulation, and this question of explanation, are best answered together. What we should aim to find is an explanation of the truth of ARC which, on the one hand, is independently plausible, and, on the other hand, supports a plausible formulation of ARC—a formulation with plausible normative implications. In the remainder of this paper, I will be exploring three alternative ways in which one might attempt to ground or explain the truth of ARC.

3. The Fairness Explanation

Let us consider some rule, R, such that if R were generally followed, then good consequences would result. R might, for example, be a rule that prohibits walking on the grass in university

courtyards. Now consider two possible worlds. In w_1, everyone follows R, and, as a result, the grass in university courtyards is healthy. In w_2, no one follows R, and, as a result, the grass in university courtyards is un healthy. Suppose, further, that in w_1, more than one person would have to walk on the grass before its health would be harmed, and in w_2, more than one person would have to refrain from walking on the grass before its health would be improved. Thus, in both cases, the health of the grass will be the same regardless of whether you walk on it. The defender of ARC will maintain that, in w_1, where everyone else follows R, you have reason to do so as well, whereas in w2, where no one follows R, you have no such reasons. How might this difference be explained, given that, in both worlds, how you act makes no difference, holding fixed the actions of others?

Perhaps the explanation is this: your failing to follow R would involve unfairness in w_1, but not w_2. In w_1, you are being a free-rider: you are benefitting from a cooperative enterprise, but you are failing to play any part in that enterprise. And that seems unfair. Could this kind of consideration explain the truth of ARC?

There is a problem with such an explanation. ARC implies that an agent *s* has reason to follow a rule R when R's being generally accepted, followed, or the like has *sufficiently good consequences*, not when it has *sufficiently good consequences for s*. This is important, because we ordinarily think that agents can be subject to a given moral requirement, even when the agent in question has never benefitted from others following this requirement. Thus, someone might be under a moral obligation to benefit the needy, even if neither she nor anyone she cares about has ever been needy, and hence even if she never benefits from the practice of benefitting the needy. Similarly, I might be under a moral obligation not to walk in the grass, even if, in virtue of a rare brain abnormality, I am incapable of responding aesthetically to foliage, and hence the health of the grass in campus courtyards doesn't affect me. But if I don't benefit from the fact that people refrain from walking on the grass, then my walking on the grass would not constitute free-riding.

Could it nonetheless be argued that I'm being unfair by failing to follow a rule that others follow and whose being followed has good consequences, when I don't benefit personally from the fact that this rule is followed? Perhaps. For in such a case, there will be some outcome that results from the rule being followed, such that everyone has *impersonal* reason to want this outcome to obtain. But if we should all want this outcome to obtain, then it might be argued that it would be fairer if we shared in the sacrifices that are made in order for this outcome to obtain. But these sacrifices consist in limiting one's actions to those that conform to rule R. And so it might be argued that its fairer if we all follow rule R.

But there remains a problem with this explanation. For it can explain why we have reason to follow a rule only if the following of this rule involves making a sacrifice. And so, at most, what this kind of explanation can support is a version of ARC that is restricted to rules following which is burdensome. However, there are plenty of moral rules that the consequentialist might

want to appeal to ARC in order to explain (insofar as they can't be explained by straightforward act-consequentialist considerations) but that are not burdensome to follow. Thus, it seems that I could have reason not to walk on the grass in w_1 even if everyone else is indifferent as to whether they walk on the grass, and hence even if refraining from walking on the grass doesn't require anyone to make sacrifices. Similarly, it seems I could have reason have reason to prioritize the near and dear even if everyone else likes prioritizing the near and dear, and hence even if no one needs to make any sacrifices in order to so prioritize.

Thus, neither the free-rider explanation, nor the kind of fairness explanation just considered, appears to support a sufficiently broad version of ARC.

4. The Agential Involvement Explanation

In this section, I will outline a theory of practical reason proposed by my colleague, Ralph Wedgwood, and I will indicate how such a theory might be used to ground ARC. The basic idea of the Agential Involvement Theory is that one's reasons to perform a given action depend on two factors: first, the *consequences* of this action, and second, the degree to which the agent would be involved in bringing about these consequences, that is, the level of *agential involvement*. Wedgwood lays out his theory as follows.

> Consider a *weighting* of the degree to which a consequence S of a course of action A instantiates an intrinsic value V by the agent's degree of agential involvement in bringing about S about (The higher the degree of agential involvement, the greater the weighting…) Call this the *agentially weighted value* of S (considered as a consequence of A with respect to V).
>
> Then we can express this theory of reasons for action in the following way: whenever there is a reason in favor of a course of action A, this reason arises from A's having a consequence S that has a *positive* agentially weighted value (of this kind); whenever there is a reason *against* A, this reason arises from A's having a consequence S that has a *negative* agentially weighted value… Other things being equal, the *higher* the positive agentially weighted value, the stronger this reason for A is; the *lower* the negative agentially weighted value, the stronger the reason *against* A.

Elsewhere he gives a more quantitative characterization of the view:

> When your act has a bad consequence, the more agentially involved you are in bringing about that consequence, the stronger the reason against the act will be … In general, the strength of the reason against the act seems to correspond to the weighted sum of the degrees of badness of each of the act's consequences—where the degree of badness of each consequence is weighted by the degree of agential involvement that the agent has in that consequence.

The same applies mutatis mutandis, he tells us, when your act has a good consequence. And so we can express the agential involvement theory of reasons for action with the following formula:

$$R(\phi) = \sum_i \left(A\left(C_i^{\phi}\right) V\left(C_i^{\phi}\right) \right) = A\left(C_1^{\phi}\right) V\left(C_1^{\phi}\right) + A\left(C_2^{\phi}\right) V\left(C_2^{\phi}\right) + \ldots$$

Here ϕ represents an arbitrary action, and $R(\phi)$ represents the balance of one's reasons for or against performing this action: where this quantity is positive, then the more positive it is the more reason one has to ϕ, and where this quantity is negative, the more negative it is, the more reason one has not to ϕ. For any value of *i*, C_i^{ϕ} represents some particular consequence of ϕ (the *i*th consequence of ϕ in our arbitrary ordering of these consequences); $V\left(C_i^{\phi}\right)$ represents the impersonal value of this consequence, and $A\left(C_i^{\phi}\right)$ represents the degree to which the agent would be agentially involved in bringing about this consequence by way of performing ϕ.

This view raises a number of questions. First, how are we to measure agential involvement? Second, how are we to individuate consequences? And third, how are we to assign values to consequences? This last question is particularly difficult, since a given action can have several consequences, and the agent may be agentially involved to differing degrees in producing these different consequences. Hence, we cannot identify the consequences of an action with the possible world that would obtain if the action were performed, nor can we read off the value of a given consequence of an action from the value of the world that would obtain were the action performed. Furthermore, where we set the zero-point on the value scale makes is practically significant, so we need a way to distinguish between good consequences and bad consequences. I won't attempt to answer these questions here, but I will say something about the constraints that are imposed on the answers if the Agential Involvement Theory is to ground ARC.

In order for this grounding to work, we need a broad conception both of what is to count as an *outcome* of an action, and of what is required for being *agentially involved* in this outcome. In particular, we need a conception on which C can count as a consequence of action ϕ performed by agent A, and A can count as agentially involved in the production of C, even if C doesn't depend counterfactually on A doing ϕ, and even if C would obtain regardless of how A were to act. Fortunately, there is independent motivation for such a broad conception of these notions. For suppose there is a firing squad consisting of ten soldiers, each of whom shoots at a convict, causing his death. It seems we would want to say that each of the ten soldiers is involved in bringing about the death of the convict, even if the shots fired by any nine of the soldiers would suffice to bring about his death, and hence even if the convict's death doesn't depend counterfactually on the action of any one of the soldiers.

Given this broad conception of agential involvement, we can see how the Agential Involvement Theory might be used to ground ARC. Suppose that the fact that people generally follow a rule requiring promise-keeping produces some good consequence, say, public trust. Just as, in the firing squad example, everyone who fires at the criminal is involved in the joint process whereby the criminal is killed, and is thus agentially involved in producing his death, so it seems that everyone who follows the promise-keeping rule is a participant in the joint process that produces public trust. Hence, they can all be regarded as agentially involved in the production of public trust. And so the Agential Involvement Theory will imply that we have a reason to follow the promise-keeping rule that derives from the fact that in doing so, we would be agentially involved in the production of public trust. And the strength of this reason will be the product of the value of public trust and the level of agential involvement we would have in producing this outcome by way of following this rule. This reason, therefore, depends crucially in two factors. First, it depends on the fact that others keep promises. For if others didn't keep promises, then public trust won't exist regardless of what we do, and so by following the promise-keeping rule we would not be agentially involved in the production of public trust. Second, our reason for keeping our promise depends on the fact that public trust is a good consequence. For, according to the Agential Involvement Theory, if public trust were a bad consequence, then the fact that our keeping our promise would involve us in the production of this consequence would provide a reason *against* promise-keeping, and if public trust were a neutral consequence, then the fact that our keeping our promise would involve us in this consequence would not provide any reason for or against promise keeping. Thus, it seems that the Agential Involvement Theory entails the existence of a reason to keep our promises of precisely the kind predicted by Actual Rule Consequentialism. And the same will apply in every case where there is a rule that is generally followed, and where this fact has a good consequence: in all such cases, the Agential Involvement Theory will imply that we have reason to follow this rule so as to be agentially involved in the production of the good consequence.

There are, however, some problems with this approach to grounding ARC. I'll begin by discussing two such problems, which are closely related. First, recall that, according to ARC, we will have reason to keep our promises if we live in a world where people generally follow a rule requiring promise keeping, but we will have no such reason if people don't follow this rule. Similarly, we'll have reason to refrain from walking on the grass if others so refrain, but not otherwise. But the Agential Involvement Theory doesn't seem to predict this asymmetry. For consider the world where no one keeps his promises. In such a world, by failing to keep my promises, won't I be participating in the joint process that produces public distrust? And so won't the Agential Involvement Theory imply that, in such a world, I have reason to avoid breaking my promise that derives from the fact that, in breaking a promise, I would be agentially involved in the production of a bad consequence? Similarly, in the world where no one refrains from walking on the grass, if I were to walk on the grass, would I not be a participant in the joint process whereby the grass is rendered unhealthy? And so won't the agential involvement theory

imply that, in such a world, I have a reason to avoid walking on the grass that derives from the fact that, were I to walk on the grass, I'd be agentially involved in producing a bad consequence?

The second problem with this approach to grounding ARC can be illustrated by way of an example. Suppose there are two buildings that need to be built: a hospital and a gas station. The hospital would require 400 workers to build, and the gas station would require only ten workers to build. Currently there are 500 workers on the scene, and each one is working on the project he or she happens to prefer. As a result, there are 491 people working on the hospital (91 more than required for its construction) and 9 people working on the gas station (one fewer than required for its construction). Then you arrive on the scene, and you must choose between working on the hospital and working on the gas station. It seems that, in this case, the best thing for you to do would be to work on the gas station. For if you do that, then both buildings will be built, whereas if you work on the hospital, then the gas station will not be built. However, the agential involvement theory seems to imply that, so long as the value of the hospital's being built exceeds the value of the gas station's being built by a sufficient margin, you will have most reason to work on the hospital. For, presumably, you will have a positive degree of agential involvement in building the hospital just in case you work on the hospital (call this degree of agential involvement *x*), and you will have a positive degree of agential involvement in building the gas station just in case you work on the gas station (call this degree of agential involvement *y*). Hence, assuming there are no other relevant consequences of either action, the strength of your reason to work on the hospital will be *x* times the value of the hospital's being built, and the strength of your reason to work on the gas station will be *y* times the value of the gas station's being built. And so, if the value of the hospital's construction sufficiently exceeds the value of the gas station's construction, you will have more reason to work on the hospital.

One might respond to this objection by denying that, in working on the hospital, one would be agentially involved in its production. After all, one would be redundant, since the hospital's construction requires only 400 workers and there are already another 491 workers working on the hospital. But this response would not appear to be available to someone who wants to use the Agential Involvement Theory to support ARC. For we want to support a version of ARC that implies that that an agent s has reason, for example, to keep her promises, even if the public goods brought about by promise-keeping would obtain regardless of whether s keeps her promises. Hence, if we are to defend such a version of ARC using the Agential Involvement Theory, we need to say that one is agentially involved in the production of the good consequences that result from the practice of promise-keeping, even if one is a "redundant" to this production, in the sense that enough other people keep their promise to bring about the goods in question.

And so anyone who intends to defend ARC on the basis of the Agential Involvement Theory appears to face a dilemma. If she claims that, in order to be agentially involved in the production of an outcome C, it suffices that one participate in a process that produces C, even if this process would produce C without one's participation, then her theory will have the

undesirable consequence that, in the hospital/gas station example, one might have most reason to work on the hospital. But if she claims that this kind of redundant participation is not sufficient for agential involvement, then her theory will not imply that we have any reason to keep our promises, if the goods secured by the public practice of promise keeping don't require our participation.

One way to solve this problem would be to draw a distinction between different ways of participating in a joint process. Perhaps the proponent of the Agential Involvement explanation could hold that, when one participates in a joint process that produces an outcome C, and when this outcome would obtain regardless of whether one participates in this process, then one counts as agentially involved in the production of C in some case, but not in all cases. In particular, perhaps she could maintain that, where the process in question consists in the following of some rule R, then by following this rule one counts as agentially involved in the production of the consequence; whereas if one's participation in the process does not consist in one's following such a rule, and if the outcome doesn't depend counterfactually on one's participation, then one doesn't count as agentially involved in the production of the outcome. In short, perhaps she could maintain that *rule-following matters to agential involvement*.

If we make this move, and if we adopt an appropriate conception of rule-following, then we can solve both the problems discussed above. Let us say that an agent *follows* a rule R, in the relevant sense, just in case the agent complies with this rule because she takes herself to be obligated to comply with this rule. And consider, once again, w_1 (where everyone follows a rule prohibiting walking on the grass) and w_2(where no-one follows such a rule, and every one walks wherever she pleases). On the current conception of agential involvement, in w_1 there will be reason to follow the rule that prohibits walking on the grass, since in doing so one will be participating in a joint process *of rule following* whereby it is brought about that the grass is healthy, and so one will be agentially involved in bringing about this outcome. But in w_2, one will have no such reason to avoid walking on the grass. Admittedly, if one were to walk on the grass, one would be participating in a process that brings it about that the grass is unhealthy. However, since the process in question is not one of joint rule following (in the sense we have defined), one's participation in this process would not constitute being agentially involved in producing the bad outcome. Thus, we can explain why there is reason to avoid walking on the grass in w_1 but not in w_2.

Similarly, we can avoid the implication that, in the hospital/gas station case, you will have most reason to work on the hospital so long as the value of its construction sufficiently exceeds the value of the gas station's construction. For those who are working on the hospital rather than on the gas station are doing so because that's the project they prefer. They aren't doing so because they believe it's morally required to work on the hospital rather than on the gas station. And so, if you were to choose to work on the hospital rather than on the gas-station, you would not thereby be participating in a joint process of rule-following leading to the construction of the hospital. And so, on our present account of agential involvement, in making this choice one

would not be agentially involved in the construction of the hospital. And so we can avoid the implication that you have most reason to work on the hospital.

Thus, it seems the Agential Involvement Theorist could offer an explanation of the truth of ARC while at the same time avoiding the two problems discussed above. However, I have two worries about this way of explaining ARC.

First, it seems *ad hoc*. The fundamental idea behind the agential involvement theory is that we want to be involved in doing good, and we want to avoid being involved in doing bad. We have a pre-theoretic sense of what it is to be involved in producing an outcome, and it is *prima facie* plausible that, in this pre-theoretic sense, we have reason to want to be involved in producing good outcomes, and we have reason to want to avoid being involved in producing bad outcomes. If we stick to this pre-theoretic sense of involvement, then whether one counts as being involved in the production of an outcome shouldn't depend on whether the process that produces this outcome is a process of joint rule-following, in the sense defined above. And if we dispense with the pre-theoretic notion of involvement, and we adopt in its place a gerrymandered, technical notion of agential involvement, then the claim that we have reason to be agentially involved in the production of good consequences, and to avoid being so involved in the production of bad consequences, loses much of its appeal.

Second, quite apart from its connection with the ARC, the Agential Involvement Theory has some problematic features, so it's unclear that it has enough independent plausibility to offer much support to ARC. Here I will focus on just one of these problematic features. One of the main motivations for the Agential Involvement Theory is that it appears to provide a good explanation of certain deontic restrictions. Consider the requirement that one not kill one innocent person in order to prevent two other innocent people from being killed. The Agential Involvement Theorist can explain this requirement as follows. Consider the following two options:

O1: You kill person A. As a result, neither person B nor person C is killed.

O2: You refrain from killing person A. As a result, persons B and C are both killed.

Suppose that for each of A, B, and C, the value of her living is 10, and the value of her dying is -10. And suppose that if you kills someone, your level of agential involvement in her dying is 10, whereas if you refrain from killing her, your level of agential involvement in her living in 1. And suppose, that, as a result of the action you take toward one person, someone else spares of kills a second person, your level of agential involvent in this second person's living or dying is 1. In this case, according to the Agential Involvement Theory, how much reason you have to choose O1 will be ten times the disvalue of As dying, plus one times the value of B living plus one times the value of C living. That is, it will be 10x(-10) + 1x10 + 1x10 = -80. And how much reason you will have to choose O2 will be 1*10 + 1*(-10) + 1*(-10) = -10. Thus, the balance of reasons will disfavor O2 less than it will disfavor O1, and so you will have most

reason to choose O2. And so the agential involvement theory can explain why you should not murder, even to prevent two murders.

The problem with this explanation, however, is that it overgeneralizes. For the Agential Involvement theory seems to imply not only that we shouldn't murder in order to prevent two murders, but more generally that we shouldn't cause one bad consequence in order to prevent two equally bad consequences from being caused by someone else. Thus, it implies that I shoudn't cause a ketchup stain on my new rug in order to prevent two ketchup stains on my rug from being caused—after all, any stain on my rug would be a bad consequence, and I'd be more agentially involved in this consequence if I cause the stain than if I simply allow others to cause the stain. But this seems wrong: surely what I should strive to do is minimize the stains on my rug, not minimize the degree to which I'm agentially involved in the production of these stains! And further, by symmetry of reasoning, the Agential Involvement Theory seems to imply that I should save one person, even if, in doing so, I would prevent two people from being saved by others. But again, this seems wrong: surely what I should do in this case is to act in such a way that the most people are saved, not act in such a way as to maximize my level of agential involvement in the saving. Of course, the Agential Involvement Theorist could adopt a measure of agential involvement on which *causing one* good consequence doesn't have a greater positive agentially weighted value than *allowing two* equally good consequences to be caused, and on which *causing one* bad consequence doesn't have a lower negative agentially weighted value than allowing two equally bad consequences to be caused. But then she'd lose the implication that we should't murder one person in order to prevent two murders. And so she'd lose the kind of explanation that motivated the Agential Involvement Theory in the first place.

5. The Transmission Principle Explanation

In this section, I will consider one last kind of explanation of ARC. This explanation will turn on a transmission principle connecting what a group of agents ought to do together (or have reason to do together) and what the members of the group ought to do (or have reason to do).

There does appear to be some connection. Consider, for example, the group consisting of Linda and Fred. If they ought to dance together, then it seems Linda ought to dance with Fred, and Fred ought to dance with Linda. Similarly, if the group of soldiers ought to roll the bolder up the hill together, then it seems that each of the soldiers in that group ought to roll the bolder up the hill with her fellow soldiers.

Before attempting to formulate the relevant transmission principle precisely, I should first say what I mean by doing some action *together*. For any type of action ϕ, and any group of agents (call it 'the group of Gs'), let us say that say that the group of Gs *ϕ together* just in case the Gs generally (though perhaps not universally) ϕ, and in so doing they are guided by the expectation that Gs generally (though perhaps not universally) ϕ. (Or perhaps: the group of Gs *ϕ together* just in case they generally ϕ guided by the expectation that they generally ϕ guided by

the expectation that …. In what follows, I will ignore this kind of embedding for ease of exposition). Now, if a group is sufficiently small, then the only way it can be true *generally* of the members of this group that they ϕ is if it is true *universally* of the members of this group that they ϕ. Thus, in the case of the group consisting of Linda and Fred, the only way it can be true generally, of the members of this group, that they dance is if Linda and Fred both dance. Hence, on our present account of acting together, the duo consisting of Linda and Fred will dance together just in case Linda and Fred both dance, and they each do so in a way that is guided by the expectation that the other dances as well. In the case of a larger group, it could be true generally that they ϕ without its being true universally that they ϕ. Thus, if there are sufficiently many soldiers in a given group, then the group of soldiers could roll the bolder up the hill together, even if one or two of the soldiers are on the sidelines twiddling their thumbs, and even if the soldiers engaged in the rolling are aware that some of the soldiers are sitting on the sidelines twiddling their thumbs.

As a first pass at a transmission principle, one might suggest the following.

T1: For action type ϕ, any agent s, and any group to which s belongs (call it "the group of Gs"), if the group of Gs ought to ϕ together, then s ought to ϕ.

But there's a problem. Suppose Smith belongs to a group of soldiers that ought to roll a bolder up a hill together. T1 implies that Smith ought to roll the bolder up the hill. But suppose all the other soldiers refuse to roll the bolder. In this case, Smith may well be unable to roll the bolder up the hill. And since 'ought' implies 'can', the implication that Smith ought to roll the bolder is unacceptable. We might solve this problem by revising our principle, as follows.

T2: For action type ϕ, any agent s, and any group to which s belongs (call it "the group of Gs"), if the group of Gs ought to ϕ together, and s can phi, then s ought to ϕ.

But again, this won't do. For suppose, once again, that Smith belongs to a group of soldiers that ought to roll a bolder up a hill together, and that all the other soldiers in the group refuse to do so. Suppose, further, that Smith *could* roll the bolder up the hill all by himself, but that if he were to do so, he'd break his back. And suppose, finally, that it isn't particularly important that the bolder be rolled up the hill. In this case, T2 implies, incorrectly, that Smith ought to roll the bolder up the hill. So we'll need to make another revision, such as the following.

T3: For action type ϕ, any agent s, and any group to which s belongs (call it "the group of Gs"), if the group of Gs ought to ϕ together, and s can ϕ together with the other Gs, then s ought to ϕ together with the Gs.

By moving from T2 to T3, we avoid the implication that Smith ought to roll the bolder up the hill when the other soldiers all refuse, for in this case, while Smith has the option of *rolling the*

bolder up the hill, he doesn't have the option of *rolling the bolder up the hill together with the other soldiers*, and so the antecedent of T3 isn't satisfied.

But there remains a problem, which is illustrated by the following case. Suppose there are 10 soldiers, including Smith. Suppose there are two things that could be rolled up the hill: a bolder and a bomb. Suppose the group of soldiers ought to roll the bolder up the hill together, but they ought not to roll the bomb up the hill together, since, if the bomb reaches the top of the hill, it will detonate, killing everyone. Suppose that the bomb and the bolder are equally big, and all 10 soldiers would be required to roll either one up the hill. As it happens, apart from Smith, the other 9 soldiers are all attempting to roll the bomb up the hill. If and only if Smith joins in, they will succeed, and the bomb will go off.

In this case, we have stipulated that the group of soldiers ought to roll the bolder up the hill together. And so it seems to follow that they ought to do the following: *roll something up the hill together*. Since the group of soldiers ought to roll something up the hill together, and since Smith is one of the soldiers, and since Smith *can* roll something up the hill together with the other soldiers, T3 implies that Smith *ought* to roll something up the hill together with the other soldiers. However, since the only thing the other soldiers are rolling up the hill together is the bomb, it would seem that the only way Smith could roll *something* up the hill together with the other soldiers would be by rolling *the bomb* up the hill with the other soldiers. And so T3 seems to imply that Smith ought to roll the bomb up the hill together with the other soldiers. And this implication is unacceptable.

I'm not sure what the best way is to solve this problem. But here's a suggestion. Let us say that the group of Gs does action ϕ together *non-derivatively* just in case they ϕ together, and there is no more specific action ψ such that they ϕ together by way of ψ-ing together. That is, there is no more specific action ψ such that the group of Gs ϕ together in virtue of the fact that they generally ψ and in so doing they are guided by the expectation they generally ψ. We can now revise our transmission principle as follows.

T4: For action type ϕ, any agent s, and any group to which s belongs (call it "the group of Gs"), if the group of Gs ought to ϕ together non-derivatively, and s can ϕ together with the other Gs, then s ought to ϕ together with the Gs.

Moving from T3 to T4 enables us to avoid the implication that Smith ought to roll something up the hill together with the other soldiers, in the case where the only thing Smith could roll up the hill together with the other soldiers is the bomb. For *rolling something up the hill together* is not something the soldiers ought to do together non-derivatively. Rather, they ought to roll *something* up the hill together by way of rolling *the bolder* up the hill together. And since it is not the case that the soldiers ought to *roll something up the hill together* non-derivatively, we can't infer from T4 that Smith ought to roll something up the hill together with the other soldiers.

T4 concerns the transmission of *oughts* from groups to group-member, or from pluralities to members of these pluralities. But if this principle is valid, then it is plausible that there should be an analogous principle concerning the transmission of *reasons* from groups or pluralities to members, as follows.

T5: For action type ϕ, any agent s, and any group to which s belongs (call it "the group of Gs"), if the group of Gs *have reason* to ϕ together non-derivatively, and s can ϕ together with the other Gs, then s *has reason* to ϕ together with the Gs.

If we accept this kind of principle, then it seems we'll be in a position to explain a version of ARC. Here's how. Let us say that a rule R has the status of a *basic convention* in a society S just in case (1) R is generally followed by the members of S, and (2) in following R the members of S are generally guided by the expectation that R is generally followed by the members of S, and (3) the members of S don't follow R together by way of doing something else together that is more specific (e.g., following some more specific rule). Now consider the version of ARC that results when, in our schematic formulation of ARC, we substitute *having the status of a convention* for X. This version of ARC (call it *Conventional ARC*) states that an agent has reason to follow a rule R whenever (i) R has the status of a convention in the society to which this agent belongs and (ii) the fact that R has this status has sufficiently good consequences.

In order to see how we can derive this version of ARC from T5, let us suppose that conditions (i) and (ii) above are both satisfied. Since (i) obtains, we know that R has the status of a convention in the society in question, and so the members of this society generally follow R, and their doing so is guided by the expectation that the other members of the society generally follow R. Thus, the members of this society will count as *following R together*. And since they don't do so by way of doing something else together that is more specific, such as following a more specific rule, they will count as *following R together non-derivatively*. And from (ii), we know that their doing so has sufficiently good consequences. But if sufficiently good consequences result from the members of this society following R together non-derivatively, then they will clearly have a reason (deriving from these good consequences) to follow R together non-derivatively. Therefore, using T5, we will be able to infer that any given member of the society has reason to follow rule R together with the other members of the society, so long as the agent in question is able to do so. But since, from (i), R has the status of a convention in the society in question, we know that the other members of the society are generally following rule R, and their doing so is guided by the right kind of expectation. Consequently, if the agent in question follows rule R, and if her doing so is likewise guided by the right kind of expectation, then she will count as following rule R together with the other members of her society. And so the agent in question will indeed be in a position to follow rule R together with the other members of her society. And so we can infer, on the basis of T5, that the agent in question *has reason* to follow rule R together with the other members of the society. And from that we can infer that the agent in question has reason to follow rule R.

That is, on the basis of T5, we can infer that a given agent s ought to follow a rule R so long as (i) R has the status of a basic convention in the society to which this agent belongs and (ii) the fact that R has this status has sufficiently good consequences. And this is precisely what Conventional ARC states. And so we can derive Conventional ARC from T5.

In order for Conventional ARC to provide practical guidance, it would need to be precisified further. The version of ARC we have derived is fairly indefinite, since it says nothing about the strength of the reasons that it posits. We have derived this principle from a fairly indefinite transmission principle, on that says nothing about the strength of the reasons transmitted from groups to group members. But if we had a more precise transmission principle, then on its basis we might be able to derive a more precise formulation of ARC. And it seems that the most plausible version of the transmission principle would state that the strength of the group-member reason grows as the strength of the group reason grows. That is, it would state that if the Gs have a reason of strength *x* to phi together non-derivatively, and if agent s, who belongs to G, is able to phi together with the other Gs, then S has a reason to do so whose strength increases as *x* increases. And, plausibly, when there is some rule such its being followed together non-derivatively by the members of a society has good consequences, then the strength of the reason that the group of society-members have to follow this rule together non-derivatively will increase as the value of these consequences increase. In other words, how much reason the group of society members have to follow this rule together non-derivatively will increase as we increase the value of this rule's having the status of a basic convention. And so our transmission principle T5 will imply that, when an agent belongs to a society in which a rule R has the status of a basic convention, and where the fact that the rule has this status has good consequences, then the better these consequences are, the stronger will be the agent's reason to follow this rule. (Of course, the value of these consequences needn't be the only factor that is relevant to the strength of the agent's reason.)

To sum up, I have presented a moral theory, Actual Rule Consequentialism, and I have discussed three ways in which one might attempt to ground this theory. I don't mean to be endorsing any of these attempts, since I have doubts about all three. I will be very interested to find out, however, what others in the group think about ARC, and about the prospects of grounding ARC on the basis of more fundamental principles.

Postscript

I no longer endorse T4 and T5. There is an obvious problem with T4, which I'm embarrassed not to have seen immediately. Suppose I belong to some group G. And suppose a Martian will reward everyone with everlasting cake and ice cream so long as the following occurs:

(a) The G's roll the bolder up the hill together;

(b) The G's do so together non-derivatively, in the sense that they don't do so together by way of doing together some more specific kind of action. (In particular, the G's don't do so by way of carrying out a strategy wherein everyone but me rolls the bolder up the hill); and

(c) I don't participate in the rolling.

However, If I participate in the rolling, they'll punish everyone with everlasting torture. In this case, the G's ought to roll the bolder up the hill together non-derivitively, and I can do so together with the other G's. And so T4 implies that I ought to do so. And yet clearly I ought not to do so.

The transmission principle I now endorse is this:

T6: For any agent *S* belonging to group *G*, any type *t*, any group action ϕ, and any individual action type ψ, if:

(a) holding fixed everything that is not under *S*'s deliberative control at *t*, *G* ought to ϕ;

(b) *G*'s ϕ-ing would require *S*'s psi-ing

(c) Whether *S* ψs is under *S*'s deliberative control at *t*

Then: *S* ought to ϕ.

I have a derivation of ACT from T6, but it's too long to include in the margin—it involves some acrobatics to avoid what appear to be the obvious obstacles to such a derivation.

Let me make one further remark about ACT: it is not meant to be a complete moral theory, let alone a complete theory of reasons for action! It is only meant to characterize an important class of moral reasons. I also think that we have direct consequence-based reasons. And some of these direct consequence-based reasons are moral. Thus, whenever one action would have a better outcome, as evaluated impersonally, than another action, I think we have some moral reason to favor the first action to the second. ACT is meant to characterize some, or perhaps all, of the moral reasons for action that we have apart from these direct consequence-based reasons.