Wouldn't it be Nice? Moral Rules and Distant Worlds

ABELARD PODGORSKI
National University of Singapore

According to a simple version of the moral theory known as *rule consequentialism*, what we ought to do is determined by which sets of rules, if followed by everyone, would make things go best. This view, however, faces a problem sometimes called the *ideal world objection*: there are rules that would be great for *everyone* to follow, but extremely poor guides to action in our world, where some people do not adhere to them. In response, recent defenders of rule consequentialism (Brandt 1992, Hooker 2000, Ridge 2006, Smith 2010, Parfit 2011) have rejected the simple version above in favor of views designed to be less idealistic, by evaluating worlds of partial adherence instead of, or in addition to, worlds of perfect adherence.

In this paper, I will argue that these attempts to fix rule consequentialism rest on a misdiagnosis. The revisions of rule consequentialism are motivated by taking the degree of *ideality* in the worlds of evaluation as the source of the problem, and consequently, they try to avoid it by evaluating worlds with more realistic levels of adherence to norms. But the ideal world objection, I will show, is only a special case of a more general and fundamental problem that faces any view that determines what we as individuals ought to do in this world by evaluating worlds that differ from the actual world in more than what is up to us.

While the degree of ideality is a flexible feature of rule consequentialism, the commitment to evaluating distant worlds is a core feature, and I aim to show that the generalized problem, which we might call the *distant world objection*, applies not only to rule consequentialism in all its forms, but to a wide range of other moral theories which share this commitment.

Rule and Act Consequentialism

Act consequentialism (AC) claims that we ought to do whatever *act* has the best consequences.[1] Rule consequentialism, on the other hand, claims that we ought to do whatever is prescribed by those *rules* that have the best consequences. But rules are abstract—it only makes sense to talk about the consequences of a rule in the context of that rule having a certain kind of relation to the world we are evaluating: some social or psychological role, or corresponding to some pattern of behavior. When a rule has some such relation to some world, I will say that it is *adhered to* in that world. We can distinguish different varieties of rule consequentialism by how they understand this adherence.

In particular, rule consequentialist views can be distinguished along two dimensions relevant for our purposes. First, views differ according to the *type* of adherence. On some views, to determine the consequences of a rule, we look at worlds where the rule is *complied with*—where to comply with a rule is simply to act in the way it dictates. On other views, we look at the consequences in worlds where the rule is *accepted,* where this involves the rule being in some way psychologically internalized by people, something which is neither guaranteed by, nor guarantees, their compliance with that rule.[2] There are, in principle, other ways rules might find expression in a world (they might be written on stone tablets, for instance, and placed in the town square), but the distinction between compliance and acceptance covers most of the views that are seriously discussed and will suffice for describing the problem for rule consequentialism I wish to develop. That the lessons drawn will generalize to other possible forms should become clear later.

A second dimension of variation is the *degree* of adherence. Once we have decided whether to understand adherence as compliance or acceptance, we can still ask whether we should evaluate a rule based on how good the world would be if that rule were complied with or accepted by *everyone,* or based on how good the world would be if that rule were complied with or accepted by some smaller proportion of the populace.

I will start by discussing the simplest view, not because it is the most plausible or, nowadays, the most common, but because it is easiest to see how the original objection applies. This view takes *compliance* as the type of adherence, and *universality* as the degree. Specifically, it claims:

> *Universal Compliance Rule Consequentialism (UCRC):* An action A is right iff it accords with a set of rules, S, such that if everyone complied with S, the consequences would be at least as good as if everyone complied with any set of rules other than S.

The Collapse Objection

Our goal is to illustrate the ideal world objection and offer a diagnosis. But first, it is important to get on the table an old objection to rule consequentialism, because the reason that this objection

1 We can formulate versions of consequentialism in terms of *actual* consequences or *expected* consequences, but this distinction will not concern us here. I will speak for simplicity's sake throughout as though we are talking of actual consequences, but the main points of our discussion will generalize to the expected view as well.

2 For discussion of compliance vs. acceptance, see e.g. Hooker (2000), pp. 75–80.

fails against the simple view is closely connected to the reason the ideal world objection succeeds, and we will see later that some ways of refining the rule consequentialist picture to avoid our new objection expose it to the old one.

UCRC is sometimes accused of collapsing into act consequentialism (Lyons 1965, Smart 1973, Gert 2005). One way to formulate the worry is as follows: Suppose everyone is complying with a rule R. Either everyone is complying with AC as well or someone is not. If someone is not, then everyone complying with the rule "Comply with R, unless doing so fails to optimize consequences, in which case comply with AC" would have better consequences. So only rules that coincide with AC can be ideal according to UCRC. So AC and UCRC are extensionally equivalent. This would be a problem, because one of the primary motivations for the rule consequentialist picture is the thought that it can avoid certain infamously counterintuitive implications of act consequentialism, such as the suggestion that one ought to break one's promises whenever the consequences of doing so outweigh those of keeping it (see e.g. Hooker 2000, pp. 17–18).

This line of thought is mistaken on at least two counts, however. First, as Gibbard (1965) and Regan (1980) show, when the effects of our actions depend on the actions of other agents, there can be more than one way for everyone to act such that they each satisfy AC, with different overall consequences. Only if everyone is performing the *best* of the sets of actions compatible with everyone following AC are they all acting according to rules that it would be best for everyone to follow. Everyone following a (particular) UCRC-ideal rule implies everyone is following AC, in other words, but everyone following AC does not imply everyone is following a UCRC-ideal rule.

More importantly for our purposes, however, the collapse objection fails even if we set aside cases of mutually dependent actions with multiple equilibria. What the argument shows is that *when everyone is complying* with some UCRC-ideal rule they are also complying with AC. That doesn't mean that *when others are not complying* with that rule, following a UCRC-ideal rule entails following AC. In fact, we will soon see that it does not. And it is precisely the deviation from AC in cases where compliance is imperfect that leads to the ideal world objection.

The Ideal World Objection

We are now in a position to see the decisive objection against UCRC. UCRC selects rules purely based on their consequences in a world of perfect compliance with those rules. As a result, it is unresponsive to the consequences of following rules when there is less than perfect compliance, and therefore blind to problems that arise precisely because in the real world, adherence to rules is always imperfect. And while this immunizes it from the collapse argument, it has disastrous implications.

Parfit (2011, pp. 312–320) illustrates this problem by considering the rule of pacifism, which instructs one never to use violence. If everyone followed the rule of pacifism, the world would be a lovely, peaceful place. So the consequences of the rule of pacifism rate highly according to UCRC. But in a world where people do *not* follow the rule of pacifism, where there are violent, homicidal people, following the rule of pacifism prevents anyone from effectively defending themselves or protecting the weak against the cruel, potentially leading to disaster. The fact that sometimes people are violent *matters*, we think, for whether it is a good idea to follow the rule of pacifism. But UCRC doesn't have the resources to take this into account.

It is true that nearby rules like "Be pacifist, unless someone is violent first, in which case defend oneself and others" also rank highly. In fact, the consequences of universal *compliance* with this rule are the same as those of universal compliance with pacifism, for if everyone complies with this rule they will be pacifist. But this simply guarantees that such a rule will do *no better* than unconditional pacifism. If universal pacifism is good enough, following the rule of pacifism in the actual world comes out as permissible, even in cases where intuitively it seems clear one ought to defend others against violence. Worse, the same is true of *any* rule of the form "Be pacifist, unless someone is violent first, in which case do X", even when X is "kill as many people as you can" (Parfit 2011, pg. 315).[3]

In general, a rule that would be best if it were universally followed may be a disaster to follow in conditions of less than perfect compliance. Parfit dubs this the *ideal world objection*. It is easy to see that this problem does not disappear if we simply switch the type of adherence from compliance to acceptance. For the consequences of universal acceptance of pacifism may likewise be extremely good. If everyone accepted pacifism, violence would be extremely rare at best, and the acceptance of pacifism has the benefit of simplicity over more complex rules. What makes pacifism such a disastrous rule to follow in the actual world is the existence of people who neither accept nor comply with it. *Universal acceptance rule consequentialism* (UARC) is blind to the moral significance of imperfect acceptance in just the way the universal compliance view is blind to the moral significance of imperfect compliance.

The Standard Diagnosis

Variants of the objection above are now widely taken to refute the simple versions of rule consequentialism. But rule consequentialism as a project is alive and kicking. Contemporary proponents have defended modified versions of the view designed in part to immunize it against this challenge. They share an account of which feature of the simple views is responsible for the trouble. UCRC and UARC are problematic, the diagnosis goes, because they tell us to evaluate rules by only looking at *ideal* worlds, worlds where there are unrealistic levels of compliance with or acceptance of those rules.

Ridge (2006), for instance, in discussing the problem, tells us that "in the real world the phenomenon of those who do not accept an ideal moral code is all too real and generates problems which call for action. So a moral theory which provides no guidance for this would to that extent be implausibly utopian" (pg. 244). Hooker (2000) says that "we must formulate rule consequentialism so as to make room for rules about situations where there is both some non-

3 As formulated, UCRC claims that if there are multiple incompatible sets of rules with equal best consequences, it is permissible to act in accordance with any of them. One might hope to avoid the problem by modifying this feature of the view. But none of the most natural modifications of this sort will help. If our view recommends an action only when *all* the best rules agree, for instance, then because the consequences of universal compliance with "follow R, unless others violate R first, which case do X" are the same for all X, the view will not recommend anything at all. If we insist instead, as Hooker (2000, pg. 32) does, that in cases of ties we defer to the rule closest to conventional morality, then if conventional morality had rules like "Be pacifist, unless someone is violent first, in which case kill as many people as you can", it would be not only permissible but *mandatory* to kill as many people as you can. And if in cases of ties we defer to whichever rule(s) would be best to *follow* in that particular circumstance, then whenever an optimal rule R is not universally complied with (that is, in all realistic cases), we will defer to "follow R, unless R is not complied with, in which case do what has the best consequences", and the view will collapse for all practical purposes into AC.

acceptance and non-compliance with them" (pg. 82). Smith describes it as a "problem generated by the fact that, while a given moral code might produce excellent effects if everyone accepted or complied with it, it may produce extremely bad effects in a real-world situation in which there is only partial compliance or acceptance with the code" (pg. 418). And this interpretation is implicit in Parfit's labeling of the problem.

Given this very natural diagnosis, a path to a better version of rule consequentialism looks open. If the problem is that UCRC and UARC determine what we ought to do by looking only at the consequences in worlds that have unrealistically utopian levels of compliance or acceptance, we can avoid it by modifying the *degree* of adherence to the rules in the worlds we are evaluating to something more realistic. And indeed, this is the strategy contemporary defenders take. In the next section I will survey these sophisticated versions of rule consequentialism, before arguing that the diagnosis is mistaken and developing a more generalized challenge.

Rule Consequentialism Refined

The first reformed view, discussed in Brandt (1992) and defended at length in Hooker (2000), is *Fixed-Rate Rule Consequentialism (FRRC)*. According to FRRC, we ought to act according to whatever rule would have the best consequences, if adhered to at some particular rate less than 100%. How this rate of acceptance is chosen is not entirely clear, but it is supposed to be low enough that the distinctive problems of partial adherence are allowed to manifest. Hooker himself recommends 90% adherence.

A second view, *Optimum-Rate Rule Consequentialism (ORRC),* proposed (though not ultimately endorsed) by Holly Smith (2010), recommends that we act according to those rules that have the best consequences at their *optimum rate*—that is, the rate of adherence at which they have better consequences than any other rate of adherence for that rule. If rule A does best at 80% adherence, and rule B does best at 95% adherence, and 80% adherence to rule A is better than 95% adherence to rule B, then rule A is ranked higher than rule B. Smith suggests that due to certain high costs associated with universal adherence, it is likely that most rules will have an optimum rate that is less than 100%.

The third and fourth responses both suggest that we look at *multiple* rates of adherence—100%, 99%, 98%, and so on. On Michael Ridge's (2006) account, one ought to act according to those rules that do the best *on average* across the different adherence rates. Call this *Average-Rate Rule Consequentialism* (ARRC).

According to the view endorsed by Parfit (2011, pg. 317), on the other hand, one ought to act according to those rules that do best at *any* level of adherence, rather than none. Call this *Every-Rate Rule Consequentialism* (ERRC).

These views all attempt to avoid the objection against UCRC and UARC by making the rankings of rules sensitive to the consequences of imperfect adherence to those rules. Provided those consequences are sufficiently bad, each of the views will avoid recommending that we act on those rules, even if they would be great to act on if they were universally complied with or accepted. Since the rule of pacifism, for example, does poorly when even a relatively small percentage of people are willing to break it, these views will rank it lower than rules whose benefits are more stable at lower levels of adherence.

However, I will argue, all of these attempts to save rule consequentialism fail. Others

have already raised a number of objections idiosyncratic to one or other of these views—the fixed-rate view, for instance, suffers from arbitrariness concerns in its choice of rate, and the average view in its choice of averaging principle. And all of the views, by appealing to rates of adherence, face difficulties stemming from the fact that there are multiple possible ways a rate of adherence below 100% can be distributed in a population. Our evaluation of 90% acceptance of the rule "If you are in the wealthiest 10%, donate half your income to charity" will be very different if the 10% who do not accept it are the wealthiest 10% than if they are the poorest 10%. Furthermore, questions remain about how to understand the behavior of those who *don't* adhere to the rules.[4]

But I will argue that these views fail for a more fundamental reason: they misjudge the depth of the problem they were built to solve. If what the original objection shows us is just that we need to make room for the kinds of problems generated by imperfect adherence, then these responses make sense. But I think it exposes something much more troubling—a vulnerability at the very core of the rule consequentialist project. In the next section, I will construct what is effectively a toolset for building variations on the so-called ideal world objection, and then show how it can be used to undermine the attempts to save rule consequentialism.

Utility Landmines

I'll begin by introducing a handful of peculiar objects that generalize relevant features of the cases underlying the ideal world objection. I will use them instead of more familiar phenomena primarily in order to abstract away from distracting features of real-life cases and make it easier to look at potential counterexamples to rule consequentialism in a structural way. Once we see how they work, parallel examples could be constructed with more ordinary materials.

A *utility landmine* is an indestructible device that is completely inert until a trigger condition particular to each landmine, and potentially specific enough to only be satisfied in a single possible world, is met. Utility landmines come in two types—*goodmines* and *badmines*. When the trigger to a goodmine is set off, a panel on top of the mine opens up, and out pours an unimaginable quantity of rainbows, kittens, orgasm-inducing radio signals, stylish automobiles, lost Beethoven symphonies, and everything else that makes life worth living, forever and ever until the end of time. When the trigger on a badmine is set off, a panel on the top of the mine opens up and out pours an unimaginable quantity of rain, mosquitoes, nausea-inducing radio signals, unnecessary movie sequels, the worst of unrecorded garage band jam sessions, and everything else that makes life miserable, and then explodes, destroying the universe.

Utility landmines are simply an abstraction of an idea from the original ideal world objection: that there can be radical upshots, positive and negative, of relatively specific degrees of adherence. The rule of pacifism, we noted, has fantastic consequences, but only when everyone follows it. This corresponds to the existence of a goodmine which is triggered only when everyone is pacifist. A rule that says "if you play an instrument, play the oboe", on the other hand, is a rule that has rather dire consequences at very high levels of adherence. This corresponds to the existence of a badmine which is triggered at high levels of adherence to this rule.

Next, let a *dud factory* be a machine that, when turned on, produces *duds:* utility landmines of both types with trigger conditions under the following constraint—the trigger

4 See Smith (2010) for a discussion of some of these problems.

conditions are in worlds too distant to be activated by any of your actions, or for your actions to place a burden on others to trigger or avoid triggering them.[5] This constraint on the dud factory corresponds to another component of the ideal world objection: that the worlds we are evaluating are not worlds individuals have the power to realize.

Next, I want to propose three intuitive constraints on the way any plausible moral theory will treat these objects.

> 1) One must not trigger a badmine if one can avoid it without doing something even worse.[6]
> 2) One must not fail to trigger a goodmine when doing so has no comparably significant moral cost.
> 3) What one ought to do does not depend in general on whether the dud factory is turned on or what it has produced.

I take each of these constraints to be grounded in confident deliverances of commonsense morality. Translated out of the science-fictional implementation into the general phenomena they represent, they tell us that (1) we shouldn't do anything disastrous if we can avoid it, (2) we shouldn't avoid doing anything wonderful without a very strong reason, and (3) certain facts about what happens only in faraway possible worlds are irrelevant to we ought to do. The intuitive bases for (1) and (2) are fairly straightforward. To get a grip on the intuitive force behind (3), which is not directly a claim about what is right or wrong but a claim about what our moral requirements *depend* on, we can ask ourselves whether information about the existence of duds seems *relevant* to our deliberation about what to do. Imagine that someone trying to decide what to do has just learned that a dud exists. Intuitively, it seems, this should not substantially change her deliberations; it would be bizarre, given her knowledge of the dud's impotence, for her to become very concerned about which dud it happens to be—whether it would have blown up in this or that merely possible world. But on a view which violates (3), this information can be decisive.

All of the refinements of rule consequentialism, I will argue, violate some or all of these three constraints, and do so for the same fundamental reason the original version of the view failed to deal satisfactorily with partial compliance. First, though, let us revisit the simpler UCRC and UARC and see why they fail desiderata (1)–(3).

The Distant World Objection

Let R be a rule—for instance, "Do a jumping jack at noon every day". Consider a world with two utility landmines. The first is a goodmine that triggers when (and only when) R is universally adhered to. The second is a badmine which is triggered whenever it is both the case that R is *not* universally adhered to and someone does a jumping jack at noon. R is not universally adhered to

5 Alexander Dietz brought the necessity of something like the second qualification to my attention. It could be the case that nothing I will do will trigger a utility mine, but only because others will pick up the slack for my behavior. Since the imposition of such a burden is plausibly morally relevant, (3) below would not be true without something that rules out the production of mines of this sort.

6 It is hard to imagine something worse than triggering a badmine, but we may make room for views on which one may trigger a badmine to avoid, say, directly torturing a child.

in this world.

R is optimal according to both UCRC and UARC. Nothing that doesn't trigger a goodmine will have better consequences than a rule that triggers a goodmine (as long as we make the goodmine good enough). On the acceptance view, R is uniquely optimal, since no incompatible rule, if accepted, triggers the goodmine. For reasons we saw earlier, on a compliance view there will be other optimal rules, like "Follow R, unless others violate R first, in which case kill everyone you can", since everyone complying with this rule entails everyone complying with R. But R will, in any case, be among the optimal rules. So both views require or allow you to follow R, and therefore to do a jumping jack at noon, in the actual world. R, however, is not universally adhered to, so this amounts to triggering a badmine and destroying the world. Since failing to do a jumping jack is not plausibly worse than destroying the world, both views violate (1). This is, of course, just a reframing of the original ideal world objection. R is a rule that would be great if everyone accepted or complied with it, but disastrous to follow in every other case.

Note that we cannot appeal, as Hooker (2000, pg. 164) does, to a claim that the ideal code will include an "avoid disaster" rule that will prevent one from performing the jumping jack. On a compliance view, a set of rules that says "Do a jumping jack at noon even if it leads to disaster; otherwise, avoid disaster" has no worse ideal consequences than one that tells us to avoid disaster unconditionally, since no jumping jack-related disasters will be triggered in a world of full compliance with the former rule. And on an acceptance view, we can simply build into the case that the goodmine is triggered only on universal acceptance of an *unconditional* requirement to perform a jumping jack at noon, giving that set of rules better consequences than any containing an "avoid disaster" exception.

A minor adjustment of the case provides a violation of (2). Instead of a badmine which is triggered whenever someone does a jumping jack at noon and R is not adhered to, suppose there is a goodmine which is triggered only if in the actual world, nobody does a jumping jack at noon today. Because R is an optimal rule, one is required or allowed, on both UCRC and UARC, to do a jumping jack at noon. But this would be to fail to trigger a goodmine in a case where there is no significant cost to doing so.

Finally, suppose that nothing you can do will make it so everyone adheres to R. Then among the utility landmines the dud factory might create is the goodmine that triggers only if R is universally adhered to. But, as we've seen, on UCRC and UARC the existence of such mines has important effects on what you ought to do, and therefore so does the output of the dud factory. So (3) is violated. Gideon Rosen (2009) gives a counterexample to Kantian contractualism with a similar structure. In his case, a gremlin who abhors consensus will wreak havoc if any rule is universally accepted. In our terms, this amounts to placing a badmine that activates on the universal acceptance of any rule. Our intuition, Rosen thinks, is that this should be morally irrelevant. This follows from the more general constraint provided by (3).

These are, in a sense, three sides of the same coin. The feature of UCRC and UARC that is responsible for all three violations is that the evaluation of the right rules is directly sensitive to utility landmines that trigger only in distant worlds which we have no hope of realizing, but *not* directly sensitive to utility landmines in the *actual* world. This, the objection shows us, is backward.

Rule Consequentialism Refined, Revisited

Understood in this way, the problem is not resolved when we introduce the refinements involving lower rates of adherence.

Consider first the fixed-rate view. While this view makes consequences of imperfect adherence relevant, it is still the case that the rules that tell us what to do in this world are determined purely by considering what the consequences are for worlds that are not within our individual power to realize, in a way that falls afoul of (1)–(3).

All we need to do is take the examples from the previous section, and replace "universally adhered to" with "90% adhered to" (or whatever the fixed rate is). There will be rules that would trigger goodmines at 90% adherence but trigger badmines if followed in the actual world. There will be actions that would trigger goodmines in the actual world but are prohibited by rules that are best at 90% adherence. And the dud factory can produce mines that are triggered at rules being 90% adhered to, with serious implications for what we ought to do.

The same is true of the optimum-rate view. Indeed, with judiciously placed goodmines triggered at full adherence, we can *guarantee* that R's optimum rate is 100% and that its consequences at that rate are as good the consequences of any competing rule at any rate. And then the problem is just the same as the one raised against UCRC and UARC.

The two views appealing to multiple rates are slightly trickier. On these views, which rules are best is determined not just by the consequences of rules in worlds that are inaccessible to us, but also the consequences in worlds where the degrees of adherence to a given rule are what they *in fact* are. So these views are sensitive to consequences in the actual world in a way the others are not. But they are not sensitive to the consequences in the actual world in the *right* way to avoid the problem.

Here it will be helpful to separate the acceptance and compliance versions of these views. Let us take acceptance first. In order to construct a parallel objection to Ridge's Average-Rate Rule Consequentialism, all we need to do is imagine a world with a goodmine that activates whenever the acceptance level of R is higher than, say, 80%. Then R's consequences are tremendously good at 80%, 81%, 82%, and so on. Our badmine, however, only activates in the actual world, where the acceptance level is (we assume) less than 80%. Those negative consequences (and all consequences in the actual world, in fact), therefore, are swamped out, in the averaging, by the consequences in the distant worlds with high levels of adherence. So the view will tell us, in the actual world, to perform the tragic jumping jack. For the same reason, it will not advise us to trigger a goodmine by violating R, and the output of the dud factory, since it affects the utility of the rules at each level of acceptance and therefore their average, comes out as important to determining what we ought to do.[7]

On a view like Parfit's Every-Rate Rule Consequentialism, on the other hand, it's not clear why we should expect there to be any set of rules which is best at *all* levels of acceptance. Indeed, by placing goodmines that trigger for incompatible rules at different levels of acceptance, we can guarantee that one does not exist.[8]

The compliance versions of these views, on the other hand, face a dilemma, depending on how they are articulated. Take an agent in a position to act. Either evaluating a rule for at least one of its levels of compliance requires looking at the consequences of the agent themselves

7 We cannot appeal to the possibility of conditional rules to get out of this mess, because we can simply postulate that the goodmines are only triggered by acceptance of *un*conditional requirements.

8 See Ridge (2009) for a related critique.

following the rule in their actual, maximally specific circumstances C, or it does not. [9]

If the consequences of the agent's own compliance in C do *not* matter for their evaluation of the rule, then the view does not avoid the problem. Just as UCRC's evaluation procedure was blind to the consequences of following a rule whenever compliance was imperfect, this iteration of the view is blind to the consequences of following a rule in C. We can set up goodmines triggered at all the worlds we use to evaluate levels of compliance with the rule "perform a jumping jack at noon", guaranteeing the highest marks according to ARRC and ERRC, set up a badmine triggered by following that rule in C and C alone, and because of this blindness, end up endorsing the cataclysmic jumping jack.

If, alternatively, the consequences of the agent's compliance in the actual world *do* matter for a rule's evaluation at some level of compliance, the view collapses into act consequentialism. To see this, let S be a set of rules that is best at every level of compliance, or best on average. Suppose S recommends that I do something other than what AC recommends in my maximally specific circumstance C. Then the rules S*, which say "Do as S recommends, except in C, do what AC recommends", would have better consequences than S at the relevant level of compliance, because complying with S and complying with S* overlap for every case except in C, where S* does better. At all other worlds and levels of compliance, S and S* are evaluated identically. Since S* does better than S at one level of compliance, and the same everywhere else, S does not have the best consequences at every level of compliance, and does not have the best average consequences across all levels. By contradiction, S cannot recommend that I do something inconsistent with AC.

This is a general dilemma for any attempt to reconstruct rule consequentialism by granting weight to the evaluation of compliance in the actual world: if we allow the effects of compliance in the actual world veto power, then we can avoid the distant world objection, but we end up with a view that threatens to collapse substantially into act consequentialism. If we do not allow the effects of compliance in the actual world such a veto power, then the influence of distant worlds becomes overbearing and we are at the mercy of distant landmines in an objectionable way.[10]

A New Diagnosis

So the variations on Rule Consequentialism offered in response to the ideal world objection do not solve the fundamental issue. The instance of the problem that everyone saw first comes from looking at distant possibilities that are ideal. So it was natural to see this as the source of the problem, and it is not surprising that the solutions developed were built with ideality in mind. But ideal worlds are merely one kind of world that it is not up to us individually to realize, and any view which determines what we ought to do by evaluating worlds it is not within our power to make real will suffer from the more general problem. These views still make what individuals

9 These are both options, since there are many ways any given level of compliance might be realized. If an agent's following a rule would entail 60% compliance with that rule, it may seem natural to use the consequences of the agent's following to evaluate 60% compliance, but this is not a necessary component of the view.

10 I am assuming that one desideratum of a rule consequentialist theory is that it avoid making the sorts of counterintuitive claims that act utilitarianism makes about things like promise-keeping. Some, like Regan (1980), might be happy with a view that is a modest extension of act consequentialism equipped to handle cases of cooperation or collective action, and so may not be concerned about the kind of collapse here.

ought to do sensitive to *impotent* utility landmines, those that will not activate no matter what we do, and potentially *in*sensitive to *potent* landmines, those which can actually be triggered by our actions. That is the heart of the distant world objection.

The distant world objection belongs to a family of complaints which suggest that rule consequentialism makes what we ought to do too dependent on things that are in some sense "far away". Arneson (2005) argues that rule consequentialist advice is too sensitive to facts about future technological innovations, and Portmore (2009) argues that it is too sensitive to the existence of alien societies on faraway planets. But the distant world objection is the one that is most general, and gets closest to the heart of the rule consequentialist project. A view which avoided Arneson and Portmore's objections by, for instance, holding fixed present technology in the worlds of evaluation, or by looking at the consequences of adherence in the agent's own society (rather than the entire universe) might be implausible for other reasons, but it would still be recognizably rule consequentialist. The idea of evaluating worlds that differ from ours in more than our own actions, however, is the very thing that distinguishes rule consequentialism in all its forms from act consequentialism.

Two Responses

One possible response to this kind of problem, similar to one credited to Parfit by Rosen (2009) is to evaluate rules only by their effects in *normal* worlds, where being normal rules out the presence of anything like utility landmines. But this sense of "normal" is obscure, and the response looks hopelessly ad hoc. We do not ignore the effects of other machines or parts of nature when we evaluate the consequences of rules, so why these? And since utility landmines and dud factories are merely improbable and not impossible, we could find out tomorrow that the *actual* world contains such devices; to ignore the complications they bring with them when moral theorizing in a world where they might actually exist seems unmotivated and dangerous. Moreover, as discussed earlier, utility landmines are simply dramatizations of real phenomena—convenient ways to represent highly sensitive benefits and costs of an extreme kind; the world is full of natural utility landmines of a milder sort, like the effects of universal pacifism. In principle, the pattern of their consequences across possible worlds could match that of landmines, even if in practice they tend to be more moderate. Without identifying some clear difference in kind between utility landmines and these more familiar phenomena, and some reason to think that versions of (1)–(3) could not simply be given in terms of those instead, this strategy does not look promising.

Another response would be to insist that rule consequentialist considerations only provide *a reason* for and against different actions, while allowing that other considerations may matter. This would be to abandon a fully rule consequentialist picture, but it would enable our view, at least, to avoid allowing us to step on badmines in the actual world, violating (1), or encouraging us to avoid stepping on goodmines in the actual world, violating (2), since we may appeal to strong independent reasons against such actions. But if rule consequentialist reasons are to have any impact at all, they must at least sometimes be the decisive factor for what to do, when other reasons are closely balanced. So as long as the dud factory can make a difference to the rule consequentialist reasons, this view would not help avoid (3). And such a view faces a challenge—if the reasons provided by rule consequentialist considerations are too strong, then they will allow us to do quite terrible things in the actual world to respect them; if the reasons are too

weak, then the rule consequentialist element of the theory comes out largely ineffectual.

In rejecting the existence of rule consequentialist reasons for acting, it is worth noting, we are not rejecting altogether the idea of what Woodard (2007) calls "pattern-based reasons", reasons that derive from features of broader patterns of action in which our own behavior plays only a part. For example, the fact that portraying a character in a novel in a particular way would be part of a broader pattern of morally objectionable stereotyping might be a good reason against that portrayal, and the fact that by driving a getaway car one participates in a collective act of robbery is a reason not to join in, even if the robbery would take place regardless. In cases like these, the patterns that give us reasons are realized (and have objectionable features) in the *actual* world, or would be realized if we acted in some way. That is, they are *close* patterns, and so the reasons they give are not affected by the output of the dud factory. Rule consequentialist reasons, by contrast, derive from features of social patterns that are *distant*, inaccessible to us, and this is what leads to vulnerability. So the distant world objection places a constraint on how we understand pattern-based reasons, bringing out the importance of a distinction between the normative relevance of close patterns and distant ones, and casting doubt on Woodard's claim that "We can have pattern-based reasons even in cases where no one else is cooperating, and the favored pattern stands no chance of being realized" (2007, pg. xii).[11]

Making a Mess on the Way Out

The threat of the distant world objection, we have seen, extends beyond the simple versions of rule consequentialism against which it was originally raised. But the problem turns out to be even more general than rule consequentialism itself. Any view which determines what we ought to do, here and now, on the basis of an evaluation of worlds that differ from ours in more than what is individually up to us looks like it will violate at least some of (1)–(3). The dud factory produces utility mines which can have an effect on such evaluations, when intuitively its presence should be morally irrelevant. And a view that makes prescriptions by scanning distant worlds is liable to miss landmines at its feet. Even if this is not a decisive objection to such views, it raises a challenge that calls out for address—either to explain why despite appealing to distant worlds in this way the view does not violate (1)–(3) or explaining why the way it does so

11 There is not space in this paper to fully assess the positive arguments for Woodard's (2007) claim , but I will note that the main cases he uses to motivate his view, such as the putative wrongness of shooting one person to prevent a second agent from killing twenty (pp. 25–40), of working for a vile industry (pg. 79), and of participating in a harmful group act whose outcome is overdetermined by the actions of others in the group (pp. 85–90) are all very naturally explained by agent relative constraints on individual action or by objectionable features of patterns that those actions would help manifest in the actual world. Given that we do not need to appeal to reasons provided by distant patterns to get the intuitive verdicts, and that such an appeal would open itself up to at least part of the distant world objection, these other explanations are preferable. One might worry that this sort of strategy will not help explain wrongness when it comes to cases of overdetermination by *omissions*, as in the case raised by David Estlund (forthcoming) of two doctors, Slice and Patch, each of whom have a crucial part to play in a surgery but decide to play golf instead, knowing that the other will not come. But on reflection it does not seem clear that Slice acts wrongly by not doing his part, since it would be futile at best and harm the patient at worst, and there are explanations of what seems morally troubling about such cases that do not appeal to the idea that anyone acted wrongly (see Portmore (forthcoming)). In general, it is much less clear that we have reason to perform positive actions when the omissions of others render them futile than that we have reason to refrain from actively participating in moral wrongdoing whose effects are overdetermined. Dietz (2016) has further discussion on this point.

is not objectionable.

There is nothing about this difficulty which is idiosyncratic to views that assess *rules*. A form of indirect virtue consequentialism that suggests what we ought to do according to what follows from *dispositions* or *motives* it would be best to have universally or widely adopted has the same issues.[12] Any disposition or motive can be promoted as ideal, given the existence of a goodmine that triggers only when it is adopted, even if acting on that disposition or motive would lead to disaster in our world.

The objection also applies to views that go even further than the views discussed so far to avoid any hint of utopianism. Conrad Johnson (1991), for instance, defends a view, *actual* rule consequentialism, that is explicitly designed to be anti-utopian. On the type of view of which his is a member, one ought to act in accordance with those actual rules that meet a standard of being minimally justified. The actual rules are minimally justified as long as the consequences of their being adhered to are better than if act consequentialism, some other rule, or no rule at all were adhered to. Julia Driver (2007) defends a roughly analogous version of virtue consequentialism according to which it is the *actual* rather than the *ideal* benefits of character traits which determine their status as virtues.

This focus on actual rather than ideal rules or dispositions, however, does not save the view. This is because to determine whether actual rules or dispositions pass moral muster, we must make a *comparison* between the consequences in our world and certain rather distant possible worlds—those where act consequentialism, or no rule at all, or some other rule were adhered to, or where other dispositions are widely adopted. Impotent utility landmines generated by the dud factory, however, can have an effect on the evaluation of consequences in those distant worlds, and therefore make a difference to what we ought to do in the same objectionable way. To put it another way, ordinary rule or virtue consequentialism tells us to compare consequences between many distant worlds. Actual rule or virtue consequentialism tells us to compare consequences between the actual world and distant worlds. But as long as any distant worlds are involved, utility mines have a chance to do their dirty work.

Finally, the objection is not limited to views that are consequentialist. Fans of Kant's universal law formulation of the categorical imperative might likewise be concerned. That formulation says that one must act on maxims that could be willed to be universal law. While the proper interpretation of this requirement is a matter of considerable dispute, plausibly it has something to do with features of distant hypothetical worlds in which the maxim *is* universal law. If something like a utility landmine could affect those features in a way that matters for what ought morally be done, then we are in the same boat. One popular interpretation, endorsed in some form by Korsgaard (1985) and O'Neill (1975), takes a maxim's universalizability to be undermined by a "practical contradiction"—by its aim being frustrated in a world where it is universalized. But a utility mine set to activate when a given maxim is universalized can affect whether the aim of that maxim is frustrated in that world. So on this interpretation the worry remains. It might be suggested that a practical contradiction only happens if a maxim's aim is *necessarily* frustrated if universalized, and therefore that the contingent existence of utility mines is irrelevant. But then the mere possibility of such mines shows us that very few maxims will fail this test, for it is easy to construct a world where a mine guarantees that the aim of a maxim is satisfied if it were universalized, even if the aim would be undermined in normal worlds. And worries about undergeneration are precisely what motivate the practical contradiction

12 For instance, the motive or virtue consequentialism of R.M. Adams (1976), when paired with a virtue-theoretic assessment of individual actions.

interpretation over alternatives that demand some sort of logical inconsistency (see Korsgaard 1985).

Contractualist theories which judge actions according to their conformity with rules that would be agreed upon are also in danger. T.M. Scanlon's formulation, for instance, suggests: "an act is wrong if its performance under the circumstances would be disallowed by any system of rules for the general regulation of behavior which no one could reasonably reject as a basis for informed, unforced general agreement" (1982, pg. 110). It is hard to see how to interpret this in a way that doesn't somehow involve evaluating, from the individual agent's perspective at least, a world where the rules *are* the basis for unforced general agreement. Utility landmines (tailored to the agent's tastes, perhaps) can have an effect on the choice-worthiness of that world. It may be possible for a contract theory to escape this if it can give an account of the reasons contracting agents have to reject rules that *doesn't* involve evaluating a world in which the rule is adhered to. But especially given that contract theories generally try to capture thoughts about the concessions people would be willing to make, *provided that* others make similar concessions, it is unclear how such an account would go.

Notably, while this affects contractualist views across both the Hobbesian (Gauthier 1998) and Kantian (Scanlon 1982) spectrum, all of which have the above feature, it does not apply to Rawls (1971), who purports to be providing a view about which political or social arrangements are just rather than how individuals ought to act in the actual world.[13] This reveals one way to involve the assessment of distant worlds in a moral theory without falling prey to the objection from utility mines: one may evaluate those worlds in order to morally assess something broader than an individual action—an institution, practice, or collective pattern of behavior, while denying that a positive assessment of these immediately transmits a permission to *individuals* to act accordingly.[14]

The ultimate lesson to draw is something that should have, on reflection, significant pull even if we never consider obscure and science-fictional scenarios—that what matters morally about our actions is *what happens here.* We cannot count on theories that judge our actions indirectly—not by consequences or features of those actions performed in the actual world, but by what happens in other possible worlds we cannot access through our choices.[15]

References

Adams, Robert M. "Motive Utilitarianism." *Journal of Philosophy* 73(14), pp. 467–481, 1976.

Arneson, Richard. "Sophisticated Rule Consequentialism: Some Simple Objections." *Philosophical Issues*, 15(1), pp. 235–251, 2005.

Brandt, Richard. *Morality, Utilitarianism, and Rights*. Cambridge University Press, 1992.

Dietz, Alexander. "What We Together Ought to Do." *Ethics* 126(4), pp. 955–982, 2016.

13 Though the Rawlsian approach may be subject to its own, distinct worries about idealization. Valentini (2012) provides a taxonomy of these concerns.

14 Dietz (2016) discusses a view of this sort about collective actions.

15 I would like to extend special thanks to Jacob Ross, Mark Schroeder, Alexander Dietz, and Joe Horton, as well as an anonymous reviewer at *Noûs*, for many conversations on the topic of this paper and invaluable feedback on earlier drafts.

Driver, Julia. *Uneasy Virtue*. Cambridge University Press, 2007.
Estlund, David. "Prime Justice." In K. Vallier and M. Weber (eds.), *Political Utopias.* Forthcoming.
Gauthier, David. "Why Contractarianism?" In J. Rachels (ed.), *Ethical Theory 2: Theories About How We Should Live.* Oxford University Press, 1998.
Gert, Bernard. *Morality: Its Nature and Justification*. Oxford University Press, 2005.
Gibbard, Allen. "Rule Utilitarianism: A Merely Illusory Alternative?" *Australasian Journal of Philosophy* 43(2), pp. 211–220, 1965.
Hooker, Brad. *Ideal Code, Real World*. Oxford University Press, 2000.
Hooker, Brad and Fletcher, Guy. "Variable vs. Fixed–Rate Rule Consequentialism." *The Philosophical Quarterly* 58(231), pp. 344–352, 2008.
Johnson, Conrad. *Moral Legislation*. Cambridge University Press, 1991.
Korsgaard, Christine. "Kant's Formula of Universal Law." *Pacific Philosophical Quarterly* 66(1–2), pp. 24–47, 1985.
Lyons, David. *Forms and Limits of Utilitarianism*. Clarendon Press, 1965.
O'Neill, Onora. *Acting on Principle*. Columbia University Press, 1975.
Parfit, Derek. *On What Matters*, *Volume One.* Oxford University Press, 2011.
Portmore, Douglas. "Rule Consequentialism and Irrelevant Others." *Utilitas* 21(3), pp. 358–376, 2009.
———. "Maximalism and Moral Harmony." *Philosophy and Phenomenological Review.* Forthcoming.
Rawls, John. *A Theory of Justice*. Harvard University Press, 1971.
Regan, Donald. *Consequentialism and Cooperation.* Oxford University Press, 1980.
Ridge, Michael. "Introducing Variable Rate Rule-Consequentialism." *The Philosophical Quarterly* 56(223), pp. 242–253, 2006.
———. "Climb Every Mountain?" *Ratio* 22(1), pp. 59–77, 2009.
Rosen, Gideon. "Might Kantian Contractualism be the Supreme Principle of Morality?" *Ratio* 22(1), pp. 78–97, 2009.
Scanlon, T.M. "Contractualism and Utilitarianism." In A. Sen and B. Williams (eds.), *Utilitarianism and Beyond*. Cambridge University Press, 1982.
Smart, J.J.C. "Utilitarianism: For and Against." Cambridge University Press, 1973.
Smith, Holly. "Measuring the Consequences of Rules." *Utilitas* 22(4), pp. 413–433, 2010.
Valentini, Laura. "Ideal vs. Non-ideal Theory." *Philosophy Compass* 7(9), pp. 654–664, 2012.
Woodard, Christopher. *Reasons, Patterns, and Cooperation*. Routledge, 2007.