Spring 1-1-2014

On Good People: A New Defense of Rule-Consequentialism

Ryan Jenkins
ryan.r.jenkins@Colorado.EDU

ON GOOD PEOPLE
A NEW DEFENSE OF RULE-CONSEQUENTIALISM

by

RYAN JENKINS

B.A. *Summa cum laude*, Florida State University, 2008

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Philosophy

2014

This dissertation entitled:

On Good People: A New Defense of Rule-consequentialism

written by Ryan Jenkins

has been approved for the Department of Philosophy

Dr. Alastair Norcross, Associate Professor

Dr. Benjamin Hale, Associate Professor

Date: July 18, 2014

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Jenkins, Ryan (Ph.D., Philosophy)

On Good People: A New Defense of Rule-consequentialism

Dissertation directed by Associate Professor Alastair Norcross

Abstract [139 words]

Rule-consequentialism is an ethical theory that does a better job than any other of justifying our moral intuitions from a single overarching principle. My dissertation defends a novel formulation of this view as an account of *a good person*.

First, I argue that a rule-consequentialist does "what a good person would do," and that if everyone were like her, the world would be as good as possible.

Second, good people sometimes do more than is required of them, and so a theory of the good person must explain supererogation, i.e. "going above and beyond the call of duty."

Third, I argue that rule-consequentialists have good reasons to embrace different moral codes for different societies.

Finally, I show that it is fruitful to understand just war theory as a set of near-absolute constraints designed to *minimize the horror of war*.

You speak unadvisedly, my friend, if you fancy that a man who is worth anything ought not to consider only one thing when he is acting: whether he does what is right or wrong, the actions of a good man or a bad man.

—Socrates, *Apology*, 28b.

Acknowledgements

Philosophy is at once an inescapably collaborative and starkly individualistic endeavor. The fact that most works cite dozens of other thinkers while themselves having officially one *author* makes this paradox clear. This work is much the same. I would hope that the reader could overlook any mistakes, infelicities, or omissions that remain in this draft. But one sort of omission would be inexcusable, and that would be to omit acknowledgement of the others whose time and advice helped me to improve this work by leaps and bounds.

First, I'd like to acknowledge the input of my committee: Alastair Norcross, Graham Oddie, Chris Heathwood, Ben Hale, and Brad Hooker. Their impact is felt in every chapter.

I'd also like to acknowledge the generosity of Leonard Kahn, whose feedback was integral throughout this work but is most noticeable in the first and third chapters and Appendix A.

The third chapter has benefitted from more feedback than any of the others, having been a pet project for a few years. It incorporates advice from Ben Eggleston, Dale E. Miller, Duncan Purves, Bill Shaw, and Jussi Suikkanen. I also received excellent feedback during a presentation at the 17th Oxford Philosophy Graduate Conference, and while there benefitted greatly from conversations with Hilary Greaves, Brad Hooker, and Luke Davies.

The fourth chapter grew from a study conducted with my advisor, Alastair Norcross. It is currently being developed into a monograph with Leonard Kahn. Ideas and influence from both of them are apparent there. I would also like to thank the audience at the International Society for Military Ethics meeting at Notre Dame in October, 2013.

I am grateful to the Department of Philosophy at the University of Colorado Boulder for awarding me a Department Dissertation Fellowship in the Fall semester of 2013, which allowed me to complete a first draft of the dissertation in time to defend in the following Spring.

And while it's apparently customary to thank one's significant other, that only makes this overdetermined. I'd finally like to thank my wife, Gina, whose seemingly inexhaustible strength and perseverance were a constant inspiration, and who has provided the emotional support without which graduate school would have been solitary, in addition to being at times nasty and brutish.

Contents

Chapter 1

The good person as a model for others

1.1 Introduction

Many of us take it as the central goal of our lives *to be good people*. Good people are not obviously *just* those that do the right thing all of the time. Nor are they obviously *just* those who accomplish certain things with their lives. And though they do not seek to be admired, their manner, their conscience, and their psychology cry out for admiration and for emulation. They are the people we should try to be *like*; they are the people we should try to *be*.

This dissertation defends a theory of the good person, drawn from these general claims. My foundational assumption, which I will not defend, is that we should try to be good people. I argue in this chapter that the good person is the person who acts such that, if everyone followed her example, the world would be as good as possible.

Ultimately, the good person, on this account, will resemble a rule-consequentialist who, in contemporary parlance, has internalized the ideal moral code. This is because she would almost always comply with the ideal moral code, the code that if followed by everyone would make the world as good as possible.

Rule-consequentialism has typically been motivated as a foil to maximizing act-consequentialism.

> Maximizing act-consequentialism. At any time, an agent is morally required to perform the best action. The *best* action is usually understood as the action, of those available to the agent, that she could expect to maximize her contribution to impartial aggregate wellbeing. If two or more actions are tied for the best, any of these actions is permissible, but the agent is required to perform one from the set of actions tied for best.

The objections to maximizing act-consequentialism are familiar. Among the allegations leveled against maximizing act-consequentialism: it can require that we perform actions we think are impermissible, such as murdering one person to redistribute their organs; it requires too much of us, for example, demanding that we donate our wealth to curb deaths from preventable diseases up to the point of marginal utility; it requires that we see our personal projects through the lens of aggregate impersonal wellbeing and thus alienates us from those pursuits that make life valuable; it is impossible to ever do *more than* is required, as what is required is the absolute best action open to us; and so on.

In response to the veritable drubbing that maximizing act-consequentialism received in the 20th century, alternative species of consequentialism multiplied and flourished. Each of them can be seen as an attempt to improve upon maximizing act-consequentialism by avoiding the objections above. These alternative species include, but are not limited to, sophisticated consequentialism, scalar consequentialism, satisficing consequentialism, systems consequentialism, effort consequentialism, and rule-consequentialism.[1]

Rule-consequentialism is a form of indirect consequentialism, whereby our individual actions are judged in accordance with some ideal moral code of rules, where that code of rules is *in turn* judged by the aggregate wellbeing we could expect to result if the code were

[1]To be clear, this is not exclusively a species-level taxonomy: sophisticated consequentialism is a species of maximizing act-consequentialism, and effort a species of scalar.

internalized by all or almost all agents. All rule-consequentialist theories accept the following two principles:[2]:

(1) An action is right if and only if it accords with an ideal code of moral rules.

(2) A code of rules is ideal only if its acceptance by some portion of the population (usually at least the overwhelming majority of agents) would maximize expected consequences.

The ideal moral code is typically described as forbidding agents from lying, stealing, and harming others or their property, because we could expect the general observance of a code forbidding such actions to maximize impartial aggregate wellbeing. Rule-consequentialism allegedly avoids all of the objections leveled against maximizing act-consequentialism above: The ideal moral code of rules is thought to cohere better with our considered moral intuitions, e.g., it would not require that we murder one person in order to redistribute their organs to others. It does not demand that we donate significant portions of our material wealth to others who are worse off, at least not to the point of marginal utility. It does not require us to see our individual projects, or any of our actions, through a lens of aggregate impersonal wellbeing. And, finally, it is possible to go above and beyond what is required by the ideal code of rules, as is explored in a later chapter here.

Rule-consequentialism can be justified in a number of ways. Some of its proponents have argued that rule-consequentialism would actually do more to increase expected impartial aggregate wellbeing than would maximizing act-consequentialism.[3] This may seem obviously false, given the formulation of act-consequentialism as requiring each agent to always perform the best possible action at a given time. From that definition, we might think it follows *analytically* that everyone's being an act-consequentialist would have better consequences than everyone's being a rule-consequentialist, since rule-consequentialism, as is well-known,

[2]I borrow this formulation from Kahn (2013).
[3]Harsanyi (1977).

can sometimes require agents to *not* maximize the good.[4] However, rule-consequentialists have offered several replies to this response. They may assert that everyone's *attempting* to be rule-consequentialists rather than act-consequentialists would be better. Second, they have suggested that we restrict our discussion only to sets of action-guiding subjective rules that are causally possible for humans. In which case, they continue, act-consequentialism describes a subjective rule it would be impossible to consciously follow; that consciously following such a rule would have deleterious effects; and that, in any event, rule-consequentialism is *definitionally* a theory of the optimific collection of habits, chosen from among those that are causally possible for humans. The success of this response hinges on whether such a restriction of our discussion is acceptable.

The parade of justifications for rule-consequentialism continues. Some have argued that rule-consequentialism is justified by an appeal to contractualism, for example, it specifies the set of moral rules to which all rational agents have the most reason to assent.[5] Others have supported rule-consequentialism as the most rational moral system for a society to adopt.[6] Others have argued that rule-consequentialism best satisfies a pluralism of theoretical desiderata, for example, that it coheres well with our considered moral views, that it identifies a fundamental, unifying principle of morality, and that it is useful in contentious or difficult moral cases.[7] Notice, of course, that these justifications are not exclusive, as shown by Parfit's Triple Theory, a version of rule-consequentialism which appeals to several of the above justifications.[8] Finally, some have argued that rule-consequentialism is justified by a process of universalization[9]; I will argue along similar lines here.

[4]In fact, we might think it follows analytically that everyone's being a maximizing act-consequentialist would have better consequences than everyone's following *any* other moral theory.

[5]Parfit (2011).

[6]Brandt (1979), Oddie (1998).

[7]Hooker (2000: especially pp. 1–31).

[8]It may be misleading to call Triple Theory a version of rule-consequentialism since it is ostensibly a combination of three theories: rule-consequentialism, contractualism, and Kantianism. But Triple Theory, plausibly, is extensionally equivalent to some formulations of rule-consequentialism and generates many of the same moral rules as would rule-consequentialism.

[9]See Hare (1981; 1989b).

1.1.1 Prospectus

Spurred especially by the recent publication of Derek Parfit's highly anticipated *On What Matters*, I believe rule-consequentialism is poised for a renaissance. Parfit's work is a significant contribution to rule-consequentialism's ongoing convalescence after the dark night brought on by the widespread belief in the success of the dilemma of collapse and incoherence. Rule-consequentialism's second act, currently underway, was prefigured by the 1990 publication of Brad Hooker's article, "Rule-Consequentialism" in *Mind*, and was revealed more fully in his subsequent book *Ideal Code, Real World*. Among other invaluable contributions, this landmark work put to rest the dilemma of collapse and incoherence for good. Rule-consequentialism has received new attention recently, both positive[10] and negative[11], and the merits of new permutations of rule-consequentialism are being discussed.[12] This dissertation contributes to—and engages directly with many of—these ongoing debates.

This dissertation is a collection of articles that articulates a new theory of rule-consequentialism, as well as arguing in favor of rule-consequentialism on other fronts. First, in this chapter, I discuss two purported dilemmas for rule-consequentialism that have been expounded by others: first, and more familiar, is the dilemma of collapse or incoherence; second is the dilemma of utopianism or arbitrariness. I then introduce a new criticism of rule-consequentialism, what I call its *fundamental motivational tension*. I give a new argument for rule-consequentialism as a theory of what a good person would do. It will turn out that we ought to be the kind of person such that, if everyone were like us, the world be as good as possible. I call this new theory *situationist rule-consequentialism*. Finally, I will show this new argument for rule-consequentialism avoids the three criticisms just discussed.

In the second chapter, I show that rule-consequentialism is the only consequentialist view that can accommodate the supererogatory in a way that is not *ad hoc*. This is because rule-

[10]Mulgan (2005), Woodard (2008).

[11]Eggleston (2007), Suikkanen (2008).

[12]See Ridge (2006); see Hooker and Fletcher (2008), Jenkins and Kahn (In draft) for responses. See also Kahn (2012), Oddie (1998) for previous discussions of relative rule-consequentialism.

consequentialism is the only consequentialist view according to which the supererogatory is a genuine moral category, and where this results from some facet of the view that has independent motivation. If we take the supererogatory seriously as a basic moral category, then rule-consequentialism gains an advantage over its consequentialist rivals.

In the third chapter, I argue that some rule-consequentialists have good reason to be moral relativists. There, I show how, for those who hold rule-consequentialism as a result of a process of reflective equilibrium or out of a concern to maximize expected wellbeing, a form of rule-consequentialism that prescribes different moral codes depending on, e.g., an agent's social group will be an attractive option. This is because allowing multiple moral codes to be in force at the same time could plausibly make some people better off while making no one worse off, satisfying both our desires to maximize expected wellbeing and cohering well with our intuitions about costless benefits. In an appendix to the third chapter, I explore the merits and demerits of *individual relative rule-consequentialism*, the most extreme version of relative rule-consequentialism, according to which *each* person ought to follow the moral code that would allow *her* to maximize her contribution to wellbeing over the course of her life.

In the fourth and final chapter, I turn to the moral constraints on warfare. I argue that the standard collection of constraints on the conduct of warfare, known as *jus in bello* ("justice during war") within the just war tradition, can be understood as a species of rule-consequentialism. This is because the rules of warfare can be described as a set of near-absolute constraints on the conduct of war, framed with reference to some overarching consequentialist goal. I suggest that the surreptitious consequentialist goal of the just war tradition is *the minimization of the horror of war*. Thus we can unify the demands of morality in warfare by appealing to a single overarching principle. Finally, I examine the implications of this new unified theory for three contentious issues from the *jus in bello* literature: supreme emergency, the doctrine of double effect, and the problem of noncompliance. In an appendix to the fourth chapter, I explore the reasons we have for preferring a

consequentialist account of the morality of war to the communitarian one that is preferred by some prominent theorists.

1.2 Collapse and incoherence

Rule-consequentialism, in David Lyons' words, is a child of two houses: consequentialism and deontology. By occupying this space in the theoretical landscape, rule-consequentialism claims some of the advantages of both genera of views, while also exposing itself to criticisms. The most important of these criticisms is the dilemma of collapse and incoherence, rule-consequentialism's most persistent and pernicious criticism. The purported dilemma is illustrated by the following case:

> Imagine a situation wherein a rule-consequentialist could clearly maximize expected wellbeing by violating one of the rules in the ideal moral code, say, the rule to keep one's promises. It seems rule-consequentialism must require one of two things of the agent: either (1) she ought to violate the rule in the name of maximizing expected wellbeing, or (2) she ought to hew to the rule even though it will fail to maximize expected wellbeing. If rule-consequentialism requires (1), then rule-consequentialism collapses into extensional equivalence with maximizing act-consequentialism.[13] If rule-consequentialism requires (2), then rule-consequentialism requires its agents to internalize an incoherent motivational set. One the one hand, a rule-consequentialist is motivated by an overarching desire to maximize expected wellbeing. On the other hand, she must at least sometimes purposefully fail to maximize expected wellbeing, and instead act in accordance with the ideal moral code.[14] Either (1) or (2) must be true if rule-

[13]See Appendix A for a fuller discussion of some forms of collapse with regard to rule-consequentialism.

[14]The distinction between the *rule-worship objection* and this version of the *objection from incoherence* is not clear, if it exists at all. Presumably, rule-worship is meant to be objectionable in virtue of its being incoherent, or representing an incoherent motivational set. In a more extreme case, some critics allege that rule-worship could result in disaster. This charge, that rule-consequentialism could allow "one-off" disasters,

consequentialism is true. Since neither (1) nor (2) is acceptable, the theory must be false.

This misguided criticism of rule-consequentialism can be traced to a central confusion about the theory's structure. Take the following statement of rule-consequentialism's criterion of rightness:

> Formulation 1. An act is right if and only if it accords with the set of rules that would make the world as good as possible if followed by everyone.

This statement of the theory is not drawn from any particular theorist's work, but is meant to be generic enough to capture the spirit of its mainstream formulations, like those of Hooker and Parfit. It is only a short step from the above formulation to the following formulation:

> Formulation 2. If it is the case that *if* everyone were to perform action X, then the world would be as good as possible, then I should perform action X (*even if I can be sure that everyone else will not perform action X*).

In this formulation, rule-consequentialism seems to phrase its commands as conditionals: "If everyone else were to X, then you should X as well." This is a natural enough rephrasing of the first formulation above—natural, but mistaken. What this second formulation implies is that the force of rule-consequentialism's moral commands is bound up inextricably in a conditional statement. Thus, the rule-consequentialist is left with two options, which I will call *the dilemma*:

1. I ought to X just in case everyone else can be counted on to X.

has been answered previously. On Hooker's formulation of rule-consequentialism, the ideal code includes a clause that allows agents to violate a rule in order to avoid disasters (2000: 98–99). The justification for this feature of the ideal code is the same as any other: a code which includes such a clause would have greater expected consequences in terms of wellbeing, broadly construed, than a code that excluded such a clause, other things being equal.

2. I ought to X even though not everyone else can be counted on to X.

Neither of these options is palatable for the rule-consequentialist. Since we can often be confident that few others will actually *X*, the first interpretation would render all of rule-consequentialism's moral demands practically moot. On the other hand, to demand that I *X* even though others will not *X*, where the rightness of *X*ing is premised on a fact about the consequences of everyone else's *X*ing, is at best very confused and at worst incoherent. It would be incoherent because it would be to think that an action is right *only when* certain conditions obtain, namely, everyone else's performing that action, but still that I ought to perform that action regardless of whether those conditions *actually* obtain. In that case, when the explicit ground of rightness is absent, how can we be confident that the action would still be right?

This line of criticism can be traced to a mistake that takes place in inferring from the first to the second formulation of rule-consequentialism above. That mistake is a misunderstanding about the nature of the conditional statement involved in rule-consequentialism's criterion of rightness. I will call this mistaken interpretation the *causal-dependence* interpretation of the conditional that is involved. The third formulation below makes this mistake more clear:

> Formulation 3. X is the action that, if performed by everyone, would make the world as good as possible. Therefore, *if and only if everyone else actually X's*, you should X.

The third formulation makes clear what I call the causal-dependence interpretation of the conditional in this generic statement of rule-consequentialism. In such a formulation, the force of rule-consequentialism's commands depends on some conditions obtaining. Conversely, when those condition do not obtain, the force of rule-consequentialism's commands melts into air.

This would be a troubling result for rule-consequentialism. Luckily, it is mistaken. Below I give a new argument for rule-consequentialism resting on an alternative understanding of the fundamental conditional claim at the core of the view. First, I will discuss two new problems of rule-consequentialism, which the new argument also solves.

1.3 Two new problems for rule-consequentialism

1.3.1 The dilemma of utopianism or arbitrariness

Recently, critics have pressed a new objection against rule-consequentialism that arises from debates over the level of internalization for the ideal code of rules.[15] Any code of rules must be designed with a particular level of internalization in mind. One way of categorizing species of rule-consequentialism is to sort them into theories that are premised on universal internalization on the one hand[16] and theories that allow for (and expect) some amount of recalcitrance in the population.[17] These codes will differ with regard to, at least and most obviously, their instructions for dealing with scoundrels, thieves, and other rule-breakers. This is because if the ideal code were universally internalized there would be no rule-breakers. Thus a theory premised on the assumption of universal internalization would not *need* to make prescriptions for dealing with noncompliance. Since the ideal moral code is meant to be internalized, and humans can only internalize codes that are so complex, we ought to capitalize on those unused psychological resources elsewhere in our theory.[18, 19]

[15] I owe much of the discussion here to Leonard Kahn in correspondence.

[16] See Harsanyi (1977), Parfit (2011).

[17] See Brandt (1992), Hooker (2000).

[18] See my third chapter for a fuller explication of the theoretical considerations arising from rule-consequentialism's sensitivity to the internalization costs of candidate moral codes.

[19] Some rule-consequentialists would object to our drawing this dichotomy between advocates of universal internalization and "everyone else." They might argue instead there is simply a question of *what* level of internalization is best. This is true, and rule-consequentialists scarcely take their preferred level of internalization to be the defining characteristic of their peculiar theory. But it happens that advocates of particular levels of internalization can be so grouped, and there is nothing spooky or suspicious in this grouping. We might just as soon discuss a scheme that divides rule-consequentialists into those who support 90% internalization, like Hooker, and "everyone else" on either side of 90%. This categorization of rule-consequentialist

Codes that are developed under the assumption of universal internalization are thus subject to the charge of *utopianism*: they are designed for a perfect world and thus leave out rules for dealing with noncompliance, free-riders, "deterrence and rehabilitation," etc. It would be folly to expect any code to enjoy universal compliance, so any code phrased in this way would be unhelpful—and possibly counterproductive as well—in the real world, i.e. *now and around here*, to borrow a phrase from Williams.

On the other hand, codes developed under the assumption of anything less than universal internalization face at least two problems. First, all levels of internalization other than 100% seem to many to be *arbitrary*. For example, Hooker develops his ideal code on the supposition that it will be internalized by "the overwhelming majority of everyone, everywhere, in each new generation" (Hooker, 2000: 32). But when pressed, Hooker cashes out "overwhelming majority" as meaning 90% of people (2000: 80–84). As Hooker admittedly pulls this number out of thin air, to critics the arbitrariness is palpable.[20] Thus, the new dilemma for rule-consequentialists of *utopianism or arbitrariness*.

It will be worthwhile to say just a bit more about arbitrariness. In order for a level of social acceptance to be non-arbitrary, it must be so for either an intrinsic or extrinsic reason. But there is no clearly no intrinsic reason why any particular level of social acceptance is more choiceworthy than any other. One might suggest that there is something special about universal, i.e. 100%, acceptance, and so it is a non-arbitrary choice. But we might also think that 100% is simply one more number. The mere fact that a code of rules would be internalized by everyone as opposed to 90% of everyone does not seem helpful when determining whether the actions it requires and forbids are the morally right actions. Moreover, in the familiar style of a sorites case, if it is arbitrary that a code would be

theories is granted for the sake of argument, since it is not obviously flawed, and seems to be just as principled a distinction as any other categorization we could make among rule-consequentialists.

[20]See also Woodard (2008: 224–225). Jamieson and Elliot (2009) and others have lodged a similar complaint against satisficing consequentialism. In response to this particular charge, Hooker and Fletcher seem unmoved by worries about arbitrariness *per se*: "...we think the objection that focusing on any one rate of internalization is arbitrary has very little force if the rules thus selected have no implausible implications" (2008).

accepted by 100% of society minus one person, then it's unclear why getting this last person to internalize the code should make a principled difference to the code's plausibility or the force of its moral requirements.

Turning to extrinsic properties does not seem any more promising. If rule-consequentialism is justified by a kind of Kantian contractualism (Parfit, 2011: 377–380), an appeal to universalizability makes some sense. But this is not how Hooker justifies his form of rule-consequentialism (2000: 12–28, 99–102). Rather, Hooker appeals to rule-consequentialism's supreme coherence with our most deeply held moral beliefs. On this count, it is not obvious that 90% social acceptance would yield more plausible rules for behavior than another level of social acceptance.[21]

In addition to this charge of arbitrariness, theorists who phrase the ideal code in terms of less than universal internalization have to contend with another objection. Suppose we agree with Hooker that the ideal code should be formulated on the supposition of internalization by 90% of people. Imagine a world (or set of circumstances) in which this code enjoys any level of internalization other than exactly 90% and suppose that the results would be significantly suboptimal. This moral code's authority is thus weakened, since it is "designed" for a world unlike the actual world and, now and around here, its results are terrible. If that were the case, the justification for this moral code would be seriously undermined.

The possibility of a world like this obtaining reveals the quixotic nature of the task of fixing an internalization level. And for any possible internalization level we choose we can duplicate this same problem: Codes phrased in terms of $X\%$ internalization may yield terrible results if anything other than $X\%$ internalization is achieved.

Michael Ridge's article, "Introducing Variable-Rate Rule-utilitarianism," was instrumental in clarifying this new dilemma for rule-consequentialists (2006). Ridge argues that choosing *any particular level* of internalization will be at best undermotivated and at worst disas-

[21]See Kahn (Unpublisheda: 14–15) for more on this point. I return a similar point later when I introduce what I call rule-consequentialism's *fundamental motivation tension*.

trous. Ridge's solution is to abandon the idea that candidate codes of rules must stand or fall based on their expected consequences at a single level of internalization. Ridge substitutes in place of these "fixed-rate" theories what he calls "variable-rate rule-utilitarianism." According to variable-rate rule-utilitarianism an action is wrong if it violates the code of rules that has the greatest expected benefit when its expected consequences are *averaged* across "every level of social acceptance, between (and including) 0% and 100%" (Ridge, 2006: 248).[22] This variable-rate theory has three advantages over the now obsoleted "fixed-rate" versions of rule-utilitarianism that had been discussed heretofore. First, it is no longer vulnerable to the problem of arbitrariness, since it jettisons its commitment to any particular level of internalization, treating all of them as equally significant. Second, variable-rate rule-utilitarianism is not utopian, since presumably moral codes that generate rules for dealing with noncompliance will fare better than codes that do not when their expected consequences are averaged across all internalization levels. Finally, generalizing from this second advantage, variable-rate rule-utilitarianism will not fare poorly at *any* particular level of internalization, e.g. 100%, 90%, etc. (Ridge, 2006: 249).

Despite offering a solution to the dilemma of utopianism or arbitrariness, Ridge's variable-rate consequentialism has its own problems.[23] First, note that Ridge's animating concern is the avoidance of arbitrariness in rule-consequentialism. But his theory falls victim to the same problem in choosing ranges of possible internalization levels for evaluating. It would be natural to evaluate codes of rules based on integer levels of internalization, for example, at 99%, 98%, 97%, and so on. Yet this choice is clearly arbitrary, since we might as well choose to evaluate that theory at 99.9% internalization, 99.8%, 99.7%, and so on, and there

[22]Ridge says we are to evaluate the expected utility of codes at all internalization levels, 0% through 100%, inclusive. But calculating a code's expected utility at a 0% internalization level would be otiose, since all codes result in exactly equal levels of expected utility if internalized by *no one*. In this case we would simply be measuring a kind of background rate of expected utility for the world. I suspect this is just an oversight on Ridge's part.

[23]These criticisms of Ridge have been developed in cooperation with Leonard Kahn, and are given in more detail in Jenkins and Kahn (In draft).

is no reason for preferring one calculation over the other. If anything, we ought to prefer as fine-grained a calculation as is practicable.[24]

There seems to be only one principled answer here and that is to calculate the expected value of each code of rules at every possible internalization level. Internalization is a personal phenomenon, and therefore the atoms of internalization, if you will, for any given code are individual people.[25] So, in a society of 100 people—and only in a society of exactly 100 people—would it be appropriate to evaluate an ideal code of rules at the most natural integer levels of internalization: 99%, 98%, and so on. In a society of 10 people, without partial internalization levels, it would make no sense to evaluate a theory at, e.g., 94% internalization. In a society of seven-odd billion, as Hooker's locution "everyone, everywhere," makes clear is his intent, running such a calculation would be practically impossible (2000: 32).

This theoretical problem with Ridge's view has been solved, but has given way to an epistemic problem. In fact, these worries seem to undo VRU's central motivation. Ridge puts forth his theory in part because of worries about our epistemic position with regard to the level of social acceptance of any particular theory. But once the theory is examined in detail it offers less epistemic certainty than before and in fact becomes unmanageably complex.[26] Rule-consequentialists have long been vexed by the epistemic problems entailed

[24]This is on the assumption that each individual 'grain' is equally epistemically likely, but we suppose this for the sake of argument because it is plausible. See Kahn (Unpublisheda: 17–18) for more on this point.

[25]This assumes that there is no *partial* or *fractional* internalization within a single person. On the contrary, we could conceive of partial levels of internalization that are determined by the likelihood that any particular person will obey the ideal moral code on any given occasion. Suppose there are two people in the world. Without partial internalization levels, the possible levels of internalization for the ideal moral code in this world are 0%, 50%, or 100% in the cases where no one, exactly one person, and both people internalize the code. However, we can also make sense of the scenario in which one of these people has the code perfectly internalized, and the other has a rather weak allegiance to the ideal code, say, a 50% chance of following it in any given circumstance. In this situation, we might say the code's level of internalization is 75%. While this is interesting, and exponentially increases the epistemic complexity of evaluating potential codes of rules, it does not matter to the argument here or to VRU more generally: we are still to average the expected utility over the range of all possible internalization levels, regardless of whether we countenance partial internalization levels among individuals.

[26]Hooker and Fletcher (2008) express a similar worry, calling Ridge's theory "too epistemologically demanding."

by the complexity of their theory, but we leave it to the reader to judge the seriousness of these additional worries.

There is a second problem for variable-rate rule-utilitarianism that arises out of the conceptual possibility of *transfinite* expected utility at some discrete level of internalization for a code. Transfinite numbers are numbers that are larger than any finite number but not themselves absolutely infinite. Suppose some code C at some level of internalization L is expected to generate a transfinite amount of utility. Since a transfinite dividend divided by any finite divisor yields a transfinite quotient, the total expected utility of C averaged across all possible internalizations would itself be transfinite. Thus, the expected utility of C at L would *swamp* the finite expected utility of *any* other code at *any* other internalization level. In this case, C would win out against all other candidate codes in a comparison of expected utility. This is problematic because transfinite expected utility might only obtain at a single discrete level of internalization, say, at 58.283% internalization, while yielding miserable results at every other level. But any negative utility at any other internalization level, as long as it were finite, would also be swamped by the positive *transfinite* expected utility of C at L.[27] Moreover, even if C were dominated in terms of expected utility by *every* other code at *every* other level of internalization, Ridge's proposal would still require the general internalization of C.

Third and finally, averaging the expected utility of a code across all possible internalization levels yields a much less reliable result than measuring the expected utility of a code at a particular internalization level. Suppose we evaluate a code at 80% internalization. In this case, it is possible that society may reach an 80% internalization level, at which point our estimate would have been correct. However, if we judge codes by their expected utility, averaged across all internalization levels, it is much more likely that our real world results

[27]Hooker and Fletcher (2008) remark similarly that, "Ridge's theory can be skewed by an anomaly" at some particular level of internalization and that "an unusually high expected utility for a code at one particular rate of internalisation could cause a high average expected utility for the code though it is poor at every other internalization rate." Hooker and Fletcher fail to appreciate the magnitude of the problem generated by the possibility of *transfinite* expected utility in particular.

would differ significantly from our estimates. Moreover, if we have any evidence at all about what level of internalization is likely for a theory, then variable-rate rule-utilitarianism requires us to ignore that evidence in our calculations. Suppose we are evaluating some group of moral codes and we are very confident they are each likely to eventually enjoy an internalization level somewhere between 80% and 100%, say, because codes of similar complexity and demandingness have enjoyed such acceptance in the past. Ridge's theory demands in this case that we evaluate each of these codes at internalization levels of 0% through 80% inter alia, thus considering evidence that is unhelpful, distracting, or could skew the final determination.

Ridge's solution, the most prominent solution suggested to the nascent dilemma of utopianism or arbitrariness, is unsatisfying. Below, I show how my new argument for rule-consequentialism succeeds where Ridge's solution fails.

1.3.2 Rule-consequentialism's fundamental motivational tension

Consider the following two claims that Hooker makes about his rule-consequentialism:

(1) "[Rule-consequentialism] does a better job than its rivals of matching and tying together our moral convictions..." (2000: 101). That is, the ideal moral code described by rule-consequentialism does a better job than any other moral code of cohering with our most deeply held moral beliefs.

(2) The ideal moral code is the one "whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of wellbeing..." (2000: 32).

The first of these is taken from Hooker's response to the charge of incoherence, which only goes through, he argues, if rule-consequentialism's principal justification is the maximization of expected wellbeing. The second of these is an excerpt of Hooker's criterion of rightness.

Hooker means to refer to the same *code* in (1) and (2). Notice, however, that whether rule-consequentialism really *does* do the best job of explaining and justifying our considered

moral intuitions as he claims in (1) depends on the code referenced in (2). Less abstractly, there's simply no guarantee that the code that it would be best for everyone to internalize really is the moral code that coheres best with our considered intuitions.

Determining the content of the ideal moral code of rules is a daunting empirical task. In fact, the code that would maximize expected wellbeing if generally internalized might wildly conflict with our deeply held moral views. It might countenance racism or slavery. After all, other consequentialists have accepted that consequentialism could justify institutionalized slavery, both if the world had turned out differently, and perhaps as it actually was.[28]

Hooker does not give us much reason to believe that (1) and (2) refer to the same code. He does claim, and is probably right, that the two codes will have significant overlap and both prohibit, for example, harming others or their property, stealing, telling lies, breaking promises, and so on. Codes with these prohibitions would both (1) cohere well with our considered moral intuitions and (2) make the world go well if internalized by the overwhelming majority of everyone, everywhere. But having significant overlap is a far cry from being coextensive. What we need is good reason to think that the code that *best* coheres with our considered moral intuitions is the same as the code that would make the world go *best*, were it internalized by the overwhelming majority of everyone, everywhere.

According to Hooker, (1) is the best argument for rule-consequentialism. But the truth of (1) is contingent on the code picked out in (2). Without a clearer picture of the code actually picked out in (2), I suggest there exists in rule-consequentialism a *fundamental motivational tension*. The version of the rule-consequentialism I develop below avoids the fundamental motivational tension because it does not rely, in this way, on making two theoretical ends meet.

[28]See Hare (1979).

1.4 The new argument for rule-consequentialism

Recall that, above, I diagnosed the dilemma of collapse an incoherence as resting on a mistake about the nature of the conditional on which the force of rule-consequentialism's commands depends. If rule-consequentialism's *categorical* commands depend for their strength on the truth of a *hypothetical*, the antecedent of which we can be fairly confident will *not* obtain, the rule-consequentialist may be in trouble. However, there is a second interpretation of the conditional that underlies rule-consequentialism's criterion of rightness. I will call this second interpretation the *universal* interpretation. The universal interpretation of the conditional underlying rule-consequentialism's criterion of rightness is perhaps best appreciated by considering the following argument for rule-consequentialism, which I will call *the new argument*, where it makes an appearance in the fourth premise:

1. We ought to try to be good people. (Assumption)
2. A good person sets the best example for humanity with her actions.
3. So, we ought to try to set the best example for humanity with our actions. (MP 1 & 2)
4. To set an example is to act while encouraging everyone else to follow your action in morally relevantly similar circumstances.
5. The best example is the one such that, if everyone else felt encouraged to follow it in morally relevantly similar circumstances, the world would be as good as possible.
6. So, to set the best example is to act in a way such that, if everyone else were encouraged to follow your example in morally relevantly similar circumstances, the world would be as good as possible. (MP 4 & 5)

∴ So,

> Situationist Rule-consequentialism (SRC). We ought to act in such a way that, if everyone were encouraged to follow our ex-

> ample in morally relevantly similar circumstances, the world would be as good as possible. (MP 3 & 6)

1.4.1 The good person as setting the best example

Premises (2) and (5) in the above call for further defense. Why should it be that the good person is one who sets the best example for others, and how are we to judge which example is best?

It is already widely recognized that morality is *universal* such that the following claim is true:

> The moral status of an action in some set of morally relevant circumstances does not depend on the identity of the agent who performs it.

This is to say that to make a moral judgment about some case is to commit ourselves to the same judgment about any other case that is identical with regard to what Hare calls its *universal* properties (1981: 21).[29] Hare rejects the identity of the individuals that occupy various roles, such as the agent and those affected by the action (see his 1981: 111), as well as the place and time of an action, as universal properties. I will say more below about precisely which properties are morally relevant for these purposes.

Notice this claim is incredibly weak, since it is compatible with many factors *about* the agent making a difference to the moral status of her action, such as her intentions, beliefs, or past history. Whether those factors can make a difference is a matter of debate among ethicists. My concern here is not to settle this debate, but merely to advert to an uncontroversial statement of what it means for morality to be universal.

One way of supporting the above claim about the universal nature of morality is to consider its negation: "The moral status of an action in some set of morally relevant circumstances *does* depend on the identity of the agent who performs it." It seems unreasonable

[29]For more on Hare's notion of universalizability, see (1965: 12), and his essays "Objective Prescriptions" (1999a: 13–14) and "Prescriptivism" in the same volume (1999b: 19–20).

to assent to this claim, thus it is unreasonable to deny the first claim about the universal nature of morality. If we are sympathetic to this negation at all, it is likely because we are confusing an agent's identity with her desires, beliefs, or other facts *about* her that are not her identity and that—as I have already said—*may* be morally relevant.[30]

I suggest that a good person is one who takes morality seriously enough that she recognizes and appreciates the above, and realizes that if she is taking morality seriously, that what she ought to do is what everyone else ought to do in morally relevantly similar circumstances. Thus *she cannot conscientiously act without endorsing a particular action in a particular set of circumstances.* This is the sense in which her action sets an example. Allow me here to quote Sartre at some length:

> ... When we say that man chooses himself, we do mean that every one of us must choose himself; but by that we also mean that in choosing for himself he chooses for all men. For in effect, of all the actions a man may take in order to create himself as he wills to be, there is not one which is not creative, at the same time, of an image of man such as he believes he ought to be... I am thus responsible for myself and for all men, and I am creating a certain image of man as I would have him to be. In fashioning myself I fashion man. (1945)

What is incumbent upon an agent next is for her to choose her action carefully such that it represents the *best* example for humanity. Where Sartre errs is in going on to argue that no ethical system can provide an answer to the question of what I ought to do *a priori*, that I must decide alone, and that the chosen action only has value in virtue of its having been chosen. In this, he analogizes making a moral decision to creating a work of art:

> ... does anyone reproach an artist, when he paints a picture, for not following rules established *a priori*? Does one ever ask what is the picture that he ought to paint? As everyone knows, there is no pre-defined picture for him to make; the

[30] I am grateful to Leonard Kahn for suggesting this move.

> artist applies himself to the composition of a picture, and the picture that ought to be made is precisely that which he will have made... It is the same upon the plane of morality... We cannot decide a priori what it is that should be done. (1945)

For Sartre, our moral judgments necessarily depend on our commitments and, as long as we are sincere in choosing our commitments (i.e. as long as we are acting in *good faith*), then our particular commitments are irreproachable. There is no overarching standard of morality[31] by which we might judge an individual's commitments. But this surely seems wrong: it cannot be that any choice having been made, even in good faith, *just was* the right choice. Even granting that our moral decisions flow necessarily from our commitments, some commitments are better than others—namely, better commitments are the ones it would be better for everyone to share. Acting in good faith is not good enough: We must not merely set *any* example in good faith, but set the *best* example.

To borrow from Hare again, when we act, we prescribe the same action for others in similar circumstances. Thus there is a range of possible options we might choose, a range of possible examples we might set, and a range of general patterns of behavior we might endorse. It seems clear that these examples can be more or less choice*worthy*. Examples must be judged by their generalized outcome, i.e. by their consequences if generally followed.[32] Hare provides a negative criterion for choosing which principles to universalize: he says we cannot assent to any principle which we would not be willing to prescribe for others in similar situations (1965: 89). But a positive criterion naturally suggests itself: Namely, that *the most choiceworthy example* is the example it would be *best* to set for others. Since we are to judge examples by their consequences if generally followed, the best example is the

[31] That is, besides the procedural, quasi-moral, existentialist standard of *authenticity*.

[32] In light of this, Hare defends a kind of preference utilitarianism: I ought to prescribe the actions that, if universalized, would maximize the satisfaction of the preferences of all involved. While I cannot defend an axiology here, my view is at least more permissive than Hare's, and allows things besides peoples' preferences to matter in the calculus.

one that would make the world go best if everyone in relevantly similar circumstances felt pressure to follow it.

Thus, if I am taking morality seriously, I realize that I am *endorsing* a particular action in a particular set of circumstances; I am setting an example for humanity. And the example I set is apt when it is the one I would most prefer that everyone feel pressure to follow. Finally, that must be the example that I could expect to result in the best consequences if followed by everyone.

We should pause briefly here to consider examples of defective agency that seem to throw this claim into doubt. Surely, there are people who act compulsively, who lament their state, and who would *not* prescribe their behavior for others. Consider, for example, someone who compulsively washes her hands every five minutes, or someone who locks and re-locks the door ten times when she leaves her house. My first response is to note that these are not characteristically moral actions, and thus they are not inconsistent with my claim about the universal nature of morality above. However, there must be some people who, even when performing more morally loaded actions, do not prescribe their action for others. This may be either because they deny the very weak claim above about the universal nature of morality, or because they are unable to act in accordance with their own moral judgments. For those who simply deny the universal nature of morality, I have already said I find their view untenable. For those who are unable to act in accordance with their moral judgments, and thus unable to prescribe that their own behavior be generalized, I will simply note that theirs is a standard case of akratic action, which threatens equally all normative views, and does not pose a special threat to the new argument for rule-consequentialism. I cannot hope to give a full discussion of akratic action here, but will only note that these people are usually labeled *defective* for good reason.

1.4.2 Is this rule-consequentialism?

Notice that the conclusion of this argument is *only* a decision procedure that tells us what we ought to do in each individual circumstance. It does not prescribe rules of thumb, or a plurality of duties, or any such thing. In this respect, it is much like the most naive formulation of act-consequentialism discussed above, which is sometimes called *direct* maximizing act-consequentialism. Thus, situationist rule-consequentialism is admittedly a species of direct consequentialism, rather than indirect consequentialism. Thus, it is not technically a species of rule-consequentialism.

However, it seems that situationist rule-consequentialism is closely approximated by the more familiar rule-consequentialism. This is because a situationist rule-consequentialist agent could rely confidently on rules of thumb that largely resemble the kind of common-sense morality enshrined in rule-consequentialism's ideal moral code. It is plausible that a situationist rule-consequentialist ought to refrain from stealing, lying, harming others or their property, etc., because in most circumstances that will be the action that it would be the best for everyone to feel encouraged to do.

While I admit that situationist rule-consequentialism is a species of direct consequentialism I also feel that the label "*situationist rule-consequentialism*" captures the theory's nature well: each situation can be seen as the basis of an independent moral rule, and there is not obviously some ideal optimific maximally compossible set of rules that should be publicized and internalized. Another apt way to christen the theory would be as "*example act-consequentialism*" where each action is judged by whether *it* sets the best possible example in the circumstances. In the end, this question does not interest me, for the reasons Hooker gives in his response to Howard-Snyder's famous article, "Rule-consequentialism is a rubber duck."[33] If the charge is that a theory's name is unilluminating or misleading, a satisfactory response is to ask, after all, "What's in a name?"

[33]See Howard-Snyder (1993) and Hooker's response (1994).

One might respond that this ambivalence points to a deeper problem with this new theory, that there is a lacuna where the most basic foundation of the theory should be. Is there no priority between acts and rules, according to situationist rule-consequentialism? Actually, the priority of *actions* in this theory is clear, and the familiar ideal moral code of rule-consequentialism becomes a kind of shorthand for the right action. In this light, the theory again resembles a kind of act-consequentialism, which helps itself to "rules of thumb" when recommending an action. There remains a notable difference between the kinds of actions act-consequentialism requires and those that situationist rule-consequentialism would require. And there is a notable *constraint* on action, according to situationist rule-consequentialism, that also seems to disqualify it as a proper species of act-consequentialism. (This constraint, as far as I know, is unprecedented in the act-consequentialist literature.) Finally, the familiarity of rule-consequentialism's structure and general commandments, and its plausible status as a proxy for this new theory, recommends that name, it seems, quite strongly.

Notice how the conditional nature of rule-consequentialism is visible in the new argument but in a totally different form than the third (mistaken) formulation in §1.2. It is clear in the new argument that the categorical moral commands of rule-consequentialism do not depend for their force on the stipulation that everyone else is *actually* acting like you. The third formulation errs by smuggling in a particular understanding of the relationship between everyone's acting a certain way and the action thus being right. As I have said, it is not that everyone's actually acting in the relevant way is necessary before I am required to act that way, thus opening the door to the dilemma above. It is true, according to the new argument, that an action is right because it's the one that, if everyone performed it, things would go best. But its rightness does not depend on everyone's actually doing it. Rather, this dependence is not causal, but rather *ostensive*—it is a claim that picks out a particular pattern of behavior as the one it is best to set as an example, thus the one that is required of me.

It is implied by second horn of the dilemma above that there would be "no point" in *X*ing if we could be confident that not everyone else will *X*. But this thought depends on the claim that rule-consequentialism's overriding motivation is the maximization of expected welfare. Because rule-consequentialism's overriding motivation is not the maximization of expected welfare, we don't need to worry about whether the pattern of behavior that we will actually obtains. An alternative motivation is to act in ways that set the best example for society; thus, to act in the way that if everyone were encouraged to act, the world would be as good as possible *even if everyone else doesn't actually act that way.*

Nor is the new argument an argument from pattern-based reasons, as in the recent work of Christopher Woodard (2008). Woodard writes as if it's *obvious* that we have reasons to bring about a part of a good pattern, in itself, even when we can be confident that the fuller pattern will not be realized. His motivating examples are activities like voting, i.e. "I have a reason to vote even if it doesn't affect the outcome of an election because my action is a part of a good pattern." Though this kind of reasoning is surely common, it's also surely controversial among philosophers. The ultimate problem with Woodard's pattern-based reasons argument for rule-consequentialism is that his motivating examples are no less compelling than the central claim of rule-consequentialism, i.e. that I have reasons that are premised on the hypothetical actions of others, even if we can be confident that others will not act in the requisite way. Students of collective action problems can admit that Woodard's motivating examples are common, but insist that they are still problematic, and problematic in the same way, and to the same extent, as rule-consequentialism's fundamental claim. Thus, Woodard relies on a premise that is no less obvious than the conclusion he is trying to prove.

1.5 How the new argument solves the new problems with rule-consequentialism

1.5.1 The new argument and the fundamental motivational tension

The new argument for situationist rule-consequentialism given above avoids rule-consequentialism's fundamental motivational tension. The fundamental motivational tension arises out of an incongruence between two claims that are important to rule-consequentialism. The first is a claim about the justification for rule-consequentialism, namely, that it does a better job than any competing theory of unifying our considered moral intuitions. The second is a claim about the content of the ideal code of rules, namely, that the ideal moral code is the one that we could expect to yield maximum consequences in terms of wellbeing, impartially considered, if internalized by everyone. Thus the ultimate justification for rule-consequentialism depends on the content of the ideal moral code.

Not all versions of rule-consequentialism separate the theory's ultimate justification from its content. For example, in my argument above, I do not motivate rule-consequentialism out of a concern for justifying and systematizing our considered intuitions, though I find it very plausible that situationist rule-consequentialism will indeed end up proscribing those actions for the reasons Hooker gives. Of course, refraining from lying, stealing, and harming others or their property as a general policy will do a better job of contributing to impartial wellbeing than would a policy that allowed all of those. But my argument for rule-consequentialism does not rest on this claim. Rather, in my argument for rule-consequentialism, both the theory's justification and its content flow from an agent's setting an example, and her rational requirement to license those patterns of behavior that it would be best for everyone to follow. There are not, in some sense, two independent motivations for situationist rule-consequentialism. In fact, the fundamental motivational tension is only a problem for versions of rule-consequentialism that separate the content of the ideal code from rule-consequentialism's justification. Whereas, for those other views whose content and justi-

fication flow from a single principle—whose justification and content are *indivisible*, in a sense—the fundamental motivational tension will not arise. My new argument above spells out one such view.

1.5.2 The new argument and the dilemma of utopianism or arbitrariness

Recall the dilemma of utopianism or arbitrariness:

1. Every version of rule-consequentialism must premise the formulation of its ideal moral code on either universal compliance or on less than universal compliance.
2. Every version of rule-consequentialism that premises the formulation of its ideal moral code on universal compliance is *utopian*: it provides no guidance in cases of noncompliance.
3. Every version of rule-consequentialism that premises the formulation of its ideal moral code on less than universal compliance is *arbitrary*: any level of internalization that is assumed will lack a theoretical justification.
4. Therefore, every version of rule-consequentialism is either utopian or arbitrary.
5. If every version of a theory is either utopian or arbitrary, then that theory must be false.
6. Therefore, rule-consequentialism must be false.

Situationist rule-consequentialism is premised on universal internalization of the ideal moral code: it requires agents to set those examples that would be best if followed by everyone. Thus, we might expect this theory to be *utopian*: since it is premised on universal compliance, it cannot contain any guidance for dealing with scoundrels. But situationist rule-consequentialism can avoid the charge of utopianism by offering guidance in cases of noncompliance, and thus show that the above argument is unsound, since its second premise is false.

In fact, a version of rule-consequentialism can both (1) be premised on universal internalization and (2) offer guidance in cases of noncompliance. Situationist rule-consequentialism is one such theory.

Imagine that a situationist rule-consequentialist encounters a scoundrel, someone who flouts the ideal moral code of rules. Recall that, according to my view, this agent ought to act such that if everyone followed her example the world would be as good as possible. An encounter with noncompliance just becomes one more situation where relying on this decision procedure will be necessary. If this agent is faced with noncompliance or free-riding, she ought to respond in the way that, if everyone felt encouraged to respond, the world would be as good as possible. And, surely, there is some fact of the matter about the best way to respond in the her circumstances. Thus, situationist rule-consequentialism is utopian in that it is premised on universal compliance, but it also guides action successfully in situations of noncompliance. In fact, all species of rule-consequentialism that are premised on universal compliance can avoid the charge of utopianism as long as they can show that their moral theory can yield a decision procedure to guide action in cases of noncompliance.

1.5.3 The new argument and the dilemma of collapse and incoherence

We might think that situationist rule-consequentialism collapses into act-consequentialism. There are two ways this might happen, depending on an how agent deliberates about a situation.

What matters here is the way in which we describe the example we are setting.[34] For example, I may tailor my particular example such that it applies to only situations that are nearly exactly like the one I am acting under, describing the situation such that I could expect it to be precisely duplicated rarely, if ever. Though I am forbidden from referencing the identities of the people involved, or the place or time or the action, there are other ways of describing a situation that are near-definite descriptors. I may reference the color

[34]Hare recognizes the same problem (1989a: 61).

of the objects in my surroundings, for example, or the precise amount of money I have at stake in some decision, or the brand of car I am driving at the moment, and so on. In every situation, I could specify an example that, if followed, would maximize expected consequences in situations exactly like this. Because we could expect this description to obtain so rarely, there would be very little difference between saying that we ought to do A in instances that are nearly exactly like this, and simply saying that we ought to do A in *this* instance. If we performed such a procedure in every scenario we encountered, we would simply be acting as act-consequentialists.

A related problem arises if we conceive of ourselves as act-consequentialists and describe the example we are setting accordingly. For example, when asked to articulate the action we are performing, we might simply respond that we are aiming to maximize our expected contribution to impartial aggregate wellbeing. After all, if everyone followed *that* example, in all situations, then the world would be as good as possible. This is a clear analogue to the collapse objection to old versions of rule-consequentialism.

Situationist rule-consequentialism needs to thread this needle: as agents, we must see our situations as a plurality of moral considerations, but not one so complex that we fall back into act-consequentialism. The new theory can avoid these related collapse objections by saying more about the proper way to describe the scenarios we are under and the examples we are setting. Both of these questions can be answered by specifying which considerations are actually morally relevant. An example will illustrate this. Imagine the following case:[35]

> Mark is on his way to a lunch date when he sees a car that has run off the road into a ditch. Slowing down, he sees an injured person slumped over the wheel. Since this is a sparsely-traveled section of road, he's confident that he is the only one who has seen this car in some time. He knows that if he pulls over to help this person, he will miss his lunch date, breaking a promise to a friend. What should he do?

[35]This case is adapted from Ross (1930).

Suppose Mark stops to help this injured person, which seems to be the obviously right thing to do. Mark can consider his action under at least two different descriptions. First, he might see himself simply as *maximizing his expected contribution to impartial aggregate welfare*. Under another description, he might see himself as *breaking a promise in order to aid a stranger*. Depending on which of these descriptions Mark adopts, it will set a different example for humanity. If he adopts the first, he would encourage everyone to be an act-consequentialist. If he adopts the second, he endorses a particular way of viewing the moral universe as a collection of duties that sometimes conflict. From this second description, Mark would be able to approximate his moral obligations by a rule-consequentialist code, for example, one that contains rules of thumb that generally enjoin him both to keep his promises and to aid strangers in need.

I believe we have several reasons for preferring the second of these descriptions to the first. Since deliberating about a moral problem is simply one more kind of action, we can ask ourselves which method of moral deliberation sets the best example for humanity. Accordingly, I should deliberate in the way it would be best to encourage everyone to deliberate. Hare says as much: that when choosing which features of a situation are relevant, we must decide whether we could prescribe the general acceptance of moral principles that pick out those features (1981: 89). Since I would be unwilling to generalize a rule that carved out an exception for the day of the week, or the place and time at which an action was performed, those considerations cannot be relevant to my deliberations. I cannot prescribe that everyone deliberate in the same way.

There are further reasons for preferring the second method of deliberating. For one, it is a more intuitive and commonsense way of viewing moral problems. It seems that the moral universe naturally confronts us as a plurality of conflicting reasons rather than the monism that act-consequentialism would have us adopt. It is simply unnatural to deliberate as an act-consequentialist. When we consider whether to break a promise, tell a lie, harm a person, etc., we consider conflicting presumptions on either side of a question. Moreover, it

seems to me that a *good person* sees someone in need, or they see a promise to be kept, etc., rather than potential hedons waiting to be actualized.

For reasons that have been mentioned by others, it is an unwise strategy to encourage everyone to act as an act-consequentialist.[36] This is not only because we often lack the relevant information about the consequences of an action and the time to deliberate carefully, but also because we should doubt our own commitment to impartiality in weighing the effects of our actions. Thus everyone's being an act-consequentialist could have disastrous consequences for the institution of promising, but also for our sense of security against theft, harm, and deception. A world of act-consequentialists could be frighteningly unpredictable, and this is sufficient reason not to encourage everyone to deliberate that way.

Finally, we ought not construct rules that are overly complex. This is because we should want to avoid collapse into act-consequentialism for the reasons just surveyed. But also because our methods of moral deliberation should be simple enough that we can have them ready to hand at a moment's notice. They should be as clear and specific as possible while also being general enough to be easily internalized. Thus we should continue to deliberate as a rule-consequentialist would have us deliberate.

Is situationist rule-consequentialism self-effacing, then? Is it somehow unpalatable to hold that some features of a situation are (objectively) morally relevant, while also holding that people *ought not* directly consider those features of a situation when deliberating? This need not be problematic: consequentialists and other moral philosophers have accepted for some time that it may not always be best to consciously deliberate about the goals of our actions.[37] Here is David Lyons:

> Suppose we could, somehow, arrive at a general criterion of moral relevance for the description of an action. Suppose also, however, that we can determine a

[36] See, for example, (Hare, 1989a: 63) and (Hare, 1989b: 188).

[37] This is clear when deliberating about our prudential reasons, as is the case with the subjective hedonist in Railton's "Alienation, Consequentialism, and the Demands of Morality," (1984). See also Norcross (1997: 391 ff.).

> criterion of relevance which is peculiarly suited to the application of a form of utilitarian generalization [i.e. "what would happen if everyone did this?"]. These two criteria need not be identical; they may conflict. If we wish to understand utilitarian generalization, to grasp its peculiar import, we should apply such a principle with a criterion of relevance which is implied by the principle itself, whether or not such a criterion agrees with our notions of 'moral relevance'. (1965: 34)

I have been noncommittal thus far about the nature of the good. I suspect it is ultimately grounded in the wellbeing of conscious creatures, but in some moods I also accept the intrinsic value of other considerations such as friendship and desert. At any rate, whatever it is that is morally relevant need not occupy an agent while she deliberates.

A critic may object that there are really two questions here, a metaethical question of what natural facts are morally relevant and a moral question about what would make the world turn out best. It may seem that I have mistaken the first question for the second: I have appealed to expected consequences to answer a question about metaethics. But this type of response has deep roots in rule-consequentialism, which grounds many of its metaethical tenets in a claim about what would result if everyone had a certain set of beliefs and motivations. Thus what is right and wrong simply is a result of some calculation regarding peoples' beliefs and motivations. Likewise, what is essential to making a moral decision is simply a result of what set of beliefs and motivations, and patterns of reasoning, would have the best results. Thus, an appeal to what is most natural and convenient, and what would tend to have the best consequences, is adequate to ground this answer. At any rate, the claim that what an agent ought to consider when deliberating is separate from the question of what is morally relevant is familiar enough by now that I doubt it bears much more elaboration.

Moving on, is situationist rule-consequentialism incoherent? It would be incoherent if it were motivated by a concern to maximize expected wellbeing, but also sometimes required

agents to fail to maximize expected wellbeing. However, situationist rule-consequentialism is not incoherent because it contains no overarching motivation to maximize expected wellbeing. Recall that the central moral motivation for a situationist rule-consequentialist is to set the best example for humanity, where the best example is the example that would make the world as good as possible if everyone felt encouraged to follow it. There is certainly an appeal to maximizing expected consequences here, but that appeal is constrained by a universalizing consideration. Thus the primary concern is to set an example, and from within the class of examples I could set, I am rationally compelled to choose the one that would make the world as good as possible.

Earlier in this chapter I admitted that situationist rule-consequentialism is a version of direct consequentialism, albeit with a novel constraint on action. Given this, a critic might spring the following rhetorical objection: "If situationist rule-consequentialism is *not* a version of rule-consequentialism, then why should we laud it for solving these problems with rule-consequentialism? Isn't it mistaken to bill this theory as a new and improved version of rule-consequentialism when it is not, strictly speaking, in the same genus? In fact, isn't the title of this entire dissertation misleading if you're defending a version of direct consequentialism?" My response is that situationist rule-consequentialism represents an appealing alternative for those who are sympathetic to rule-consequentialism. Though a version of direct consequentialism, situationist rule-consequentialism is, as far as I can tell, nearly completely coextensive with a standard version of rule-consequentialism. Thus its commands boast the same intuitive plausibility.[38] Moreover, it supplies the theoretical resources for solving the problems with rule-consequentialism just discussed. Thus, if we are interested in retaining rule-consequentialism's intuitive appeal but uneasy about its theoretical problems, situationist rule-consequentialism represents a clearly superior alternative.

[38]Where situationist rule-consequentialism diverges from standard rule-consequentialism, for example, in dealing with noncompliance, situationist rule-consequentialism's differences commend it.

Situationist rule-consequentialism, as a fledgling moral theory, capitalizes on some of the most appealing insights in the history of ethics. Namely, it captures the Sartrian notion that when I choose, I choose for all mankind; that my actions implicitly endorse a pattern of behavior for humanity; that I say, in a real sense, "This is how everyone should be."

Situationist rule-consequentialism is admittedly complex, and it is unclear whether it is practically empirically evaluable. In this way, it enjoys good company with nearly every other moral theory worth considering: it seems the best we can do is to yield plausible answers to difficult questions, while certainty will always elude us. Uncertainty is a fixture of the human condition, of course, and we should not be surprised if it obscures our moral deliberations. The best we may hope for is that quantum of clarity found in the knowledge that with our actions, we do our best to present the noblest example of humanity.

Chapter 2

Rule-consequentialism and the supererogatory

Supererogatory actions are actions that, colloquially, go "above and beyond" the call of duty.[1] They are actions that are neither required, wrong, nor merely permissible. Typically they are actions that go beyond what is required, count as "overscrubscribing" to moral duty, or involve some significant personal sacrifice or risk that is admirable while not required.

The supererogatory is a fundamental moral category that we acknowledge in our daily experience of the moral world. We appeal to the supererogatory when we describe actions like a soldier falling on a grenade to save his comrades, or a doctor traveling a great distance to help needy patients in the developing world. It would be unsatisfying for a moral theory to label these actions as obligatory or merely permissible—in fact, it should count against that theory's plausibility. I suggest that any theory that deflates the supererogatory, provides an error theory, or relegates it to a second-order existence is also worse off in a respect for its inability to approximate our moral experience.

[1]The supererogatory itself is a perennially puzzling notion in ethics. For a discussion of its various species, see Feinberg (1961). For a seminal exploration of the supererogatory, see Urmson (1958).

The paradox of the supererogatory[2] is that the concept of supererogatory actions highlights an uneasy tension between two fundamentally different moral notions: the right and the good. Moral philosophers usually speak as if there are three categories of moral appraisal:

(1) Actions that are good to perform and bad not to perform (*obligatory* or *required* actions).

(2) Actions that are neither bad to perform nor bad *not* to perform (*merely permissible* actions).

(3) Actions that are bad to perform and not good to perform (*wrong*, *forbidden*, or *prohibited* actions).

Supererogatory actions do not fit into this threefold classification, as they are typically understood as actions that are *good to perform but* not *bad not to perform*. How could some actions be good without *also* generating a moral duty? We normally think that there is some essential connection between our axiological judgments of value and deontic judgments of obligation, even if the order of explanation constitutes one of the fundamental disagreements between consequentialists and non-consequentialists. But the supererogatory makes both camps uneasy. While deontological theories have had some trouble accommodating the supererogatory Baron (1987), Guevara (1999), Haydar (2002), the trouble is much more pronounced among consequentialists. This difference seems traceable to the inclusion of *options* in typical deontological approaches, which seem to be necessary for the possibility of genuine supererogation. If more than one option is open to an agent, and one of the options involves personal sacrifice in a way that is admirable, or involves oversubscribing to a moral theory, then that action is plausibly supererogatory. Consequentialist theories typically deny that there are options, unless two or more actions are tied for the best, in which case one of them cannot *pay out* any more than the other. Here I will argue that rule-consequentialism is alone among consequentialist theories in being able to accommodate the supererogatory.

[2]For discussion of this paradox, see Dreier (2004), Horgan and Timmons (2010).

2.1 Consequentialism and supererogation

Consequentialist theories have traditionally had trouble accommodating the supererogatory. Consider the most popular consequentialist theories on offer. According to *maximizing act-consequentialism*, an act is right just in case its expected consequences are at least as good as those of any other action available to the agent. According to maximizing act-consequentialism, once an agent has performed the action that was required of her then, by definition, there is nothing *more* she could have done. Since there is no difference between the best (or most admirable) action and the action that is morally required, maximizing act-consequentialism just has no room for the supererogatory.

Since, according to maximizing act-consequentialism, there are no supererogatory actions, some have objected that this theory is *too demanding*. In response to this and other objections, Peter Railton has developed a new species of consequentialism called *sophisticated consequentialism*. If sophisticated consequentialism has a single raison d'être, it is to avoid the *paradox of hedonism*: It is widely believed that if a person directly aims at achieving some goal, she will often frustrate her own efforts. Thus the best way of achieving some goal is often *not* by aiming at it directly. This is the paradox of hedonism. Railton contemplates an analogous worry for maximizing act-consequentialism: that aiming at the production of welfare could result in less welfare being produced than if something else were the direct aim Railton (1984).

But sophisticated consequentialism has other advantages. Maximizing act-consequentialism is often alleged to be *too demanding*. We see shades of this objection in the previous paragraph, where we are reminded that maximizing act-consequentialists are on every occasion required to produce as much good as possible. Thus, there is nothing they can do that goes "above and beyond the call of duty." In addition to ruling out the supererogatory, a maximizing act-consequentialist may also be overcome by a feeling of alienation: alienation from her projects, from her loved ones, from the things that she takes to give her life meaning. Maximizing act-consequentialism seems to require agents to

see her projects and relationships through the lens of aggregate impartial welfare: I ought to abandon my projects, friends, and loved ones just in case they are not optimific. Many find this demand callous and unacceptable. Sophisticated consequentialism is a species of maximizing act-consequentialism that, according to Railton and others, makes a rigorous consequentialism compatible with a more normal life. It does this by requiring an agent to internalize a set of dispositions that could be expected to maximize her contribution to impartial welfare over the course of her life, and these dispositions are allegedly compatible with close personal relationships, personal projects, etc.

It may seem at first that sophisticated consequentialism allows us to live normal lives. From this we might infer that it rescues the supererogatory, since we can go above and beyond the call of duty in our normal lives. But looking beneath the surface reveals that those actions that constitute a normal life are, strictly speaking, wrong actions according to sophisticated consequentialism. This is because sophisticated consequentialism is ultimately a species of maximizing act-consequentialism. Because sophisticated consequentialism is a species of maximizing act-consequentialism, the right action in any particular instance is the action that contributes the maximum expected amount of wellbeing to the world, even though this is frequently not the action supported by the ideal set of dispositions. In this somewhat bizarre fashion, even the scupulous sophisticated consequentialist must accept that she frequently, or perhaps always, performs the wrong action. She merely denies that this is blameworthy. So, supererogation is impossible according to sophisticated consequentialism, despite first appearances. Nor can the sophisticated consequentialist say that there is some better set of dispositions that could be internalized, i.e. a supererogatory *set* of dispositions, since the set of dispositions that she is required to internalize is already the best possible one. So much, then, for sophisticated consequentialism.

Satisficing consequentialism was developed in response to the same charge of demandingness against maximizing act-consequentialism. On this view, an action is right if it produces at least as much expected good as some baseline in some particular circumstance, i.e. just

in case it is "good enough."[3] For example, suppose in some circumstance this baseline is 50 net units of expected wellbeing. Performing any action that produces 50 net units of expected wellbeing would discharge an agent's moral obligation. Any action that results in *more* than 50 net units of expected wellbeing is supererogatory since the agent would have done *more than* was required of her. Yet satisficing consequentialism entails moral judgments that many theorists find wildly implausible.[4] For example, suppose I am in a situation where producing 50 units of welfare would be good enough according satisficing consequentialism. But suppose I could also produce 100 units of welfare with no greater cost to myself. In this case, if I fail to produce the better outcome, especially given that it is *no more demanding* than an outcome that is "good enough," I have done something wrong.

This charge underscores a flaw in the very structure of satisficing consequentialism: that any line drawn between *morally required* and *merely good enough* is bound to be arbitrary and unsatisfactory. Surely, there is some moral difference between producing 50 units of wellbeing in the example and producing 49 units, but this small difference cannot ground the difference between an action's being permissible and being wrong.[5] Satisficing consequentialism thus accommodates the supererogatory but only at the cost of theoretical plausibility since its accommodation is objectionably *ad hoc*.

One response is to move to incorporate other morally relevant factors into the determination of what counts as "good enough." For example, what is good enough need not be merely producing some level of expected wellbeing, but producing some level of wellbeing *given* a host of other factors, including the effort and personal risk involved. To return to our above example, the satisficing consequentialist might respond that it would be wrong for an agent to fail to produce the better of the two outcomes *given* that she could produce

[3]For a defense of satisficing consequentialism, see Slote (1985). For a formulation of satisficing consequentialism as a decision procedure rather than a criterion of rightness, see Simon (1966).

[4]See Hooker (2006: 239), which draws on Mulgan (2005).

[5]Norcross (2005b) agrees that there is no principled reason for specifying a 'cut-off point' between what is forbidden and what is "good enough" in some circumstance. Jamieson and Elliot (2009: 244) note that maximizing consequentialism at least has a principled reason for specifying the demands on agents: they ought to produce what they expect to be the *best* world.

either outcome with equal effort. However, satisficing's central flaw of arbitrariness is still present, since what counts as "good enough" for any particular level of effort, personal risk, etc., will still be arbitrary. Supposing I am required to expend N units of effort to discharge my moral obligations, we might as well as why it is N and not $N-1$.

Finally, according to *scalar consequentialism*,[6] there are strictly speaking no actions that are morally required (or forbidden). Since there are no actions that are morally required, there are no actions that go *beyond* what is morally required. Scalar consequentialism holds that we have moral reason to perform some actions, and indeed more reason to perform some actions rather than others. However, proponents of scalar consequentialism deny that moral reasons for or against acting generate moral obligations and prohibitions. It thus dispatches with our traditional moral categories of requirements and prohibitions. Scalar consequentialists may hold that the categories of rightness and wrongness—and hence also the supererogatory—are of practical significance, even if they are not fundamental moral categories.[7] Thus, the category of the supererogatory contains "actions that are considerably better than what would be expected of a reasonably decent person in the circumstances" Norcross (2005b). Scalar consequentialism thus enables us to make true judgments that some actions are supererogatory, but those judgments reduce to statements about our mere expectations of an agent's behavior in the relevant circumstance. This is troubling because it mistakes the order of explanation: our expectations of an agent's behavior in some circumstance are only justified if, in fact, supererogation is a real (mind-independent) property. Scalar consequentialism risks building on sand if it defers to our expectations about others' behavior with no further grounding. But that grounding can only come in the form of a prior property of actions, i.e. that they are supererogatory. Mere expectations are not morally important. For scalar consequentialists, the supererogatory is ultimately a linguistic con-

[6] See Norcross (2005b) for the canonical statement of this view.

[7] See Norcross (2005a) on the implications of this linguistic thesis for scalar and other consequentialist theories.

vention. If we wish to maintain that the supererogatory is instead as a fundamental moral category, then scalar consequentialism must be rejected.

The problem is apparent. Maximizing act-consequentialism and scalar consequentialism cannot accommodate the supererogatory. Satisficing consequentialism can but only by incurring a substantial theoretical cost. What is required of consequentialist theories is (i) to allow room for the supererogatory as a genuine, basic moral category, and (ii) to do so in a way that is neither arbitrary nor *ad hoc*, i.e. in virtue of some feature of the theory that has independent justification. There are two ways to understand this second criterion, and satisficing consequentialism fails under both understandings. First, we might mean a theory's bare ability to accommodate the supererogatory in a way that has independent justification. Satisficing consequentialists might think they can satisfy the criterion under this understanding: As Norcross (2005b) points out, satisficing consequentialism was developed in response to three worries about maximizing act consequentialism. Those worries were that "(1) it requires too much sacrifice of agents, (2) leaves inadequate room for moral freedom, and (3) does not allow for supererogation." I have argued that satisficing consequentialism avoids this third problem, but satisficing consequentialists might point out that avoiding (1) and (2) count equally well as motivations for specifying some baseline of moral obligation. Thus, the theory's accommodation of the supererogatory does have some independent motivation. However, it should be noted—as Norcross does note—that these three worries are simply facets of the general problem of *demandingness*: the claim that morality offers us *options*, thereby affording us a degree of personal freedom, just is the claim that we are not always morally required to do what would be best, which in turn implies the claim that there are supererogatory actions. Thus, to argue along these lines that satisficing consequentialism boasts three independent motivations for setting a baseline would be triple counting. But we might understand the second criterion differently: perhaps it is not the bare ability of a moral theory to account for the supererogatory, but rather its way of determining whether any particular action is supererogatory. Above I argued that any particular baseline for judg-

ing an action "good enough" will inevitably be arbitrary. Thus, satisficing consequentialism illuminates both understandings of this second criterion in its varied ways of failing it.

None of the consequentialist theories discussed above can satisfy both of these requirements. Below I show that rule-consequentialism can meet both of them successfully.

2.2 Rule-consequentialism and supererogation

Rule-consequentialism refers to a family of ethical views that are committed to the following two principles[8]:

(1) An action is right if and only if it accords with an ideal code of moral rules.

(2) A code of rules is ideal only if its acceptance by some portion of the population (usually at least the overwhelming majority of agents) would maximize expected consequences.

Rule-consequentialist theories differ with regard to their understanding of *acceptance*[9] in the above criterion and the specific *level*[10]. Formulated as above, rule-consequentialist theories are also neutral with regard to their understanding of what makes the acceptance of codes better or worse, i.e. their account of the *good*. None of these questions need concern us here. Rather I will give an account of how rule-consequentialists, uniquely among all consequentialists, and regardless of how they answer these questions of formulation, can give a satisfactory account of the supererogatory.

The claim that rule-consequentialism can accommodate the supererogatory is immediately confronted with a problem: if actions are already judged by some optimal standard of

[8]I borrow this formulation from Kahn (2013).

[9]For some, acceptance of a rule just is *successful compliance* with a rule, i.e. always obeying a rule. More popular nowadays is the view that acceptance of a rule implies the formation of a more complicated multi-track disposition to *want* to obey the rule *because it is the rule*, to encourage others to obey the rule and discourage them from breaking it, to feel good when following a rule and guilty when breaking it, etc.

[10]For theorists who formulate the ideal code of rules in terms of universal acceptance, see Brandt (1979), Harsanyi (1977), Parfit (2011). For theorists who formulate the ideal code of rules in terms of less than universal, but still overwhelming, acceptance, see Hooker (1990; 2000). For discussion of the problems facing either approach, see Jenkins and Kahn (In draft), Kahn (Unpublishedb), Ridge (2006).

morality—in this case, their coherence with the *ideal moral code*—then how can any action go *beyond* what is required? No action could be *better* than what the ideal code required: if an action differs from what the ideal code requires, that action must violate the ideal code, and must therefore be wrong. Hence, rule-consequentialism seems to squeeze the supererogatory out of the picture in the same way that maximizing act-consequentialism does. I will call this the obvious problem regarding supererogation for rule-consequentialism, or *the obvious problem* for short. This analogy is perhaps clearer between rule-consequentialism and sophisticated consequentialism, since the two theories require agents to act from an ideal set of dispositions. One difference between these two theories—their criteria of rightness—is crucial for rescuing rule-consequentialism from the same fate as sophisticated consequentialism.[11]

One response is this: According to rule-consequentialism, an action is supererogatory any time it would have *better expected consequences* than the action required by the ideal code of rules. But supererogatory actions, at the very least, are praiseworthy actions, and praising agents who violate the ideal moral code in the interest of maximizing expected consequences threatens to reintroduce the specter of *collapse*. This is surely dangerous ground for rule-consequentialists, who have labored diligently in the past to avoid falling into this particular trap.[12]

Both a closer examination of the supererogatory and a closer examination of the structure of rule-consequentialism are in order. Only after both will we see how this proposed solution fails and how rule-consequentialism avoids the obvious problem.

This proposed solution errs in assuming that the best understanding of supererogatory acts is that they contribute to the general welfare to some degree that is over and above what is required by agents. But this is mistaken, for supererogatory acts are more obviously

[11]These theories are contrasted in greater detail in Appendix A.

[12]Strictly speaking, this objection differs from the traditional collapse objection. That objection says that rule-consequentialism collapses into equivalence with act consequentialism in the actions it requires and forbids. The objection I press here does not regard rightness but rather supererogation. Still, this objection bears a striking resemblance to the collapse objection, since its point is that rule-consequentialists should be loath to endorse a moral code that would result in practical equivalence with act consequentialism.

those that involve enduring significant personal sacrifice or risk,[13] in ways that are morally admirable. Take for example the following case, inspired by a case found in Urmson's canonical article (1958):

> Doctor. Imagine a medical doctor who works in the United States. One day, she hears of an epidemic threatening a village in sub-Saharan Africa. Moved by the plight of the people there, the doctor suspends her practice in America and flies to Africa to help treat the victims of the epidemic.

We intuit that this doctor has done something supererogatory.[14]

In investigating whether rule-consequentialism can accommodate the supererogatory, we should ask whether the theory could coherently account for our most deeply held beliefs about that category of actions. Since our intuitions pick out admirable personal sacrifice or risk as the essential features of supererogatory actions, we should ask whether rule-consequentialism can make sense of *those* features. Here is how it can:

On rule-consequentialism, supererogatory acts are those that

(1) Involve significant personal sacrifice or exposure to significant personal risk, where *significant* is understood as going beyond what is required by the ideal moral code,

(2) Do so in a way that is admirable, and

(3) Do not violate any rule in the ideal moral code.

[13]It has been suggested to me that risk could be subsumed under sacrifice since, in *risking* her fortune or wellbeing, a person sacrifices her *security*, or lowers her expected future wellbeing.

[14]Likewise, take the following story from my own experience: While walking to the bar one night through a residential neighborhood, I passed a car whose dome light was on. If left on all night, the car's battery would likely be dead in the morning. I approached the house nearest where the car was parked and notified the woman who answered the door. But this was not her car. I looked around and it was not immediately obvious in which house the car's owner lived. Moreover, there were two or three apartment complexes immediately nearby, and I was not about to knock on 20 doors on my way to the bar. So, I walked on. My action seemed supererogatory and the woman with whom I spoke, while it was not her car, praised me more than once for doing something "awfully nice" (presumably a folk term for *supererogatory*). This is plausibly because the *expected* consequences of my action were high, even if the consequences that obtained were lower than if I had not knocked on the door, since it inconvenienced both me and the woman with whom I spoke.

The first criterion above captures the conceptual core of the supererogatory: doing more than is required or, literally, *paying out above* what is required. The second criterion is meant to preempt worries about the wrong kind of sacrifice, such as pointless or even vicious sacrifice. As Feinberg (1961) points out, mere sacrifice cannot be enough to make an action supererogatory. This example of his is excellent, if dated: Suppose someone approaches me on the corner and asks me for a match. In return, I give him two matches. Though this involved overpaying in a way that leaves me worse off than is strictly required, this action is not supererogatory, only odd. Moreover, agents can sacrifice in ways that are not admirable, but are pointless. For example, suppose that in some circumstance it is admirable for me to visit my aunt in the hospital—suppose she lives several cities away. I could sacrifice by biking to the hospital rather than driving. But such sacrifice does not make my action supererogatory, only wasteful. It is true that I endure greater sacrifice than I would have if I had driven my car, but the sacrifice is totally frivolous and not admirable. Thus, choosing to bike rather than to drive does not make my action supererogatory. Feinberg also contemplates "supererogatory villains," if such a label is even coherent, who are *extra* villainous (Feinberg, 1961: 282). Both considerations show that in order for my action to be supererogatory, the sacrifice involved must be positively admirable. The third criterion above is entailed by the second since an action is admirable only if it does not violate the ideal code of rules. For clarity that entailment is made explicit in the third criterion.

For rule-consequentialists, the heuristic for determining whether a candidate rule belongs in the ideal moral code is to ask, "What would the world be like if everyone *felt compelled* to follow this rule?" (Hooker, 2000: 5). The account of the supererogatory given above fits neatly into this characterization. Recall that supererogatory acts are those that are admirable yet not required. According to the ideal moral code, some amount of personal sacrifice is required, for example, in donating to aid organizations in the developing world. Universalizing a rule that would require greater sacrifice than this could very clearly have counterproductive

consequences.[15] For that reason, according to rule-consequentialism, agents ought not feel as though they are *required* to endure greater personal sacrifice for the sake of those who are worse off. On the other hand, great personal sacrifice is not strictly forbidden by the ideal code of rules, and so is permissible. This is perhaps the clearest example of "oversubscribing" to one's moral duty. Finally, supererogatory acts are admirable only if the agent assumes great personal sacrifice or risk in the pursuit of *otherwise* morally permissible goals. Thus, rule-consequentialism can accommodate supererogatory actions as those that go beyond the ideal code in terms of admirable personal sacrifice without explicitly violating any of its rules.

2.2.1 Rule-consequentialism and options

Above we noted that maximizing act-consequentialism simply has no room for the supererogatory. This is because there are no actions that are *better than* what is required—and if some action is better than what is required in terms of personal sacrifice, then it will likely be suboptimal and so forbidden. Maximizing act-consequentialism requires the agent to perform the action with the best expected consequences and thus does not provide *options* for an agent.[16]

Rule-consequentialism differs from maximizing act-consequentialism because it accords options to the agent. Consider the following two candidate rules:

(R1) Each person who is at least moderately well-off is required to donate *exactly* 10% of her income to charity.

[15] See Carson (1991), Hooker (1991), and especially Hooker (2000: 152–158). Witness also Peter Singer's diminishing demands on agents for giving to charity. In his (1972), those demands only ceased when agents had reached the point of diminishing marginal utility. In his (1982: 246) Singer analogized charity to the Christian requirement to tithe, and so suggested agents donate 10% of their income to charity. Singer's most recent and ongoing work on the topic, including his (2010) and its accompanying website, `www.thelifeyoucansave.com`, asks that we donate only 1% of our income to charity.

[16] See McNaughton and Rawling (1998), which argues that offering options is perhaps one of the distinctive features of deontology (along with agent-relative reasons and constraints). Strictly speaking, there is one kind of case in which maximizing act-consequentialism allows an agent to choose between one or more options, and that is a case wherein two or more options are *tied* for the best in terms of expected consequences. As discussed above, though, this does not really allow room for the supererogatory.

(R2) Each person who is at least moderately well-off is required to donate *at least* 10% of her income to charity.

We may call the first rule *definite* and the second rule *open-ended*. If the ideal moral code contains only definite rules, then the obvious problem with rule-consequentialism will obtain: It will be impossible for an agent to do more than what the ideal moral code requires without at the same time violating the ideal moral code, and so there would be no supererogation. If—and only if—the ideal moral code contains at least some open-ended rules will it be possible for an agent to act in a way that is supererogatory. (This is, again, only if they do not in the process violate some other rule in the ideal moral code, for example by undermining their abilities to discharge their financial obligations to their dependents.)

It is nearly certain that the ideal moral code will contain at least some rules that are open-ended rather than definite. A rule regarding charity, listed as (R2) immediately above, is an open-ended candidate moral rule that we can be very confident will be included in the ideal moral code. This is for two reasons. First, we could expect the world to turn out better if everyone internalized a moral code containing (R2) rather than (R1). This is because a code containing (R1) rather than (R2) explicitly forbids an agent from donating more than 10% of her income to charity, while (R2) leaves this open. Thus it is implausible that a society of agents having internalized (R2) would donate *less* to charity than a society having internalized (R1). Second, (R2) does not obviously have a higher psychological cost than (R1), since the two rules make identical minimum demands on the agent and leave it open to her whether to donate more than 10% of her income. Thus, it is extremely implausible that the internalization of a moral code containing (R2) would have *worse* expected consequences than a code containing (R1) instead; in a sense, there is nowhere for the expected consequences to go but up. Now we can appreciate why the obvious problem rests on a mistake: it assumes that all the rules in the ideal moral code are definite rather than open-ended. But this is almost certainly wrong.

2.2.2 Relative rule-consequentialism and the supererogatory

A new species of rule-consequentialism called *relative rule-consequentialism* has attracted attention recently.[17] According to relative rule-consequentialism,

(1) An action is wrong for an agent to perform if it is forbidden by the code of rules to which she is subject.

(2) An agent is subject to a code of rules if and only if her being subject to that code of rules is in accordance with the ideal arrangement of all codes of rules.

(3) An arrangement of codes of rules is the ideal arrangement of all codes of rules if and only if it is the arrangement that would have maximum expected consequences in terms of wellbeing.

A full explication of relative rule-consequentialism is a project for another time.[18] For such an explication to be satisfying, it would at least have to give some general idea of the boundary lines between codes. Are codes to be relativized according to differences in intellectual capacities, innate altruistic motivation, national boundaries, "social groups," or some other determinant (or combination thereof)?

According to relative rule-consequentialism, different moral codes may contain different obligations regarding charity, for example. If this is true, then which actions count as supererogatory will also be determined by each individual code. It will be easier for an agent to perform a supererogatory act if the code that governs her behavior is less demanding to begin with. So, it may be that agents with, for example, lesser innate psychological capacities or more self-interested dispositions could perform actions that would be supererogatory *for them* while not being so *for others* who fall under the purview of a different ideal moral code.

[17] While relative rule-consequentialism has been entertained since at least Brandt (1979), it has only recently attracted serious attention (see Kahn (2012), Oddie (1998), and the third chapter of this dissertation).

[18] See the third chapter of this dissertation, which argues that rule-consequentialists of certain stripes have good reason to relativize the ideal moral code to societies.

Whether this outcome is objectionable likely depends on the contours of the various ideal codes in force at once. We may think it is perfectly reasonable that someone with a strong natural reluctance to sacrificing for others ought to be praised when they donate a middling amount to charity since, for them, the action was performed only after significant psychological exertion. Thus, psychological or mental characteristics might constitute some of the less objectionable grounds for distinguishing between the moral obligations of different groups of agents.

We may think it is "reasonable" to praise a more stingy person because it was "only fair." But in another sense, a deep understanding of someone's character might lead us to say that their action *really was* more praiseworthy. Suppose we consider our obligation to give to charity in absolute numbers. Then, it would be reasonable to praise the person working three minimum wage jobs for giving $100 to charity much more than we would praise a millionaire for doing the same. This is because it's actually much more difficult for the former. We accept this kind of relativity without question in these cases, and this is why we are tempted to move to a percentage-based rule regarding charity to begin with. From here it's a fairly short step to saying that someone who finds it naturally hard to behave well should receive more praise. If we shape our institutions of praise and blame in part by reference to consequentialist goals, then it might be best to praise people who find it difficult to do the right thing more emphatically than we praise those who would find the same act easier, since we could expect this general policy to produce more wellbeing.

On the other hand, any suggestion that we draw this line based on national boundaries—similarly, ethnic or racial boundaries—is likely to encounter significant resistance. Ultimately, whether relative rule-consequentialism's handling of the supererogatory is objectionable depends on features of the theory that will have to remain unsettled for now.

Strictly speaking, we act in accordance with the ideal moral code any time we do the bare minimum, i.e. the least amount required of us. Taking examples (R1) and (R2) above,

this means we would discharge our duty of charity any time we donate (merely) 10% of our income. And many agents are like this; they are *minimally decent rule-consequentialists*. But we ought to distinguish these agents from those who routinely go above and beyond, who are *saintly rule-consequentialists*, and who distinguish themselves even in a world of good people.

Chapter 3

Rule-consequentialism and two forms of moral relativism

Rule-consequentialism determines the rightness of actions by reference to some ideal code of rules which is generally taken to apply to all (or almost all) agents. This is true in its canonical formulations, for example, Brandt (1979), Harsanyi (1977), Hooker (2000), and Parfit (2011). Relativized forms of rule consequentialism, which specify different moral rules for different social groups, have been considered before[1], yet they have not received serious attention until recently.[2] Here I argue that, depending on their theoretical motivations, some rule-consequentialists have very good reasons to be relativists. Namely, rule-consequentialists who find compelling the theory's coherence with our considered moral intuitions and are moved by consequentialist considerations ought to support a scheme of multiple relativized moral codes.

I take Hooker's view of rule-consequentialism to be a paradigm of contemporary rule-consequentialism, though what I say applies to any rule consequentialists who share his motivations. Moreover, my discussion exposes a startling inconsistency in Hooker's view,

[1] See for example Brandt (1979: 194) and Hooker (2000: 189).

[2] See especially Kahn (2012). See Oddie (1998) for an earlier discussion.

which already countenances one kind of moral relativism, namely one according to which the rightness of actions can change over time, but vigorously rejects the kind being defended here.

3.1 Four Worlds

All rule-consequentialist theories accept the following two principles:[3]:

(1) An action is right if and only if it accords with an ideal code of moral rules.

(2) A code of rules is ideal only if its acceptance by some portion of the population (usually at least the overwhelming majority of agents) would have the best expected consequences.

A necessary component of any plausible candidate for the ideal moral code is a rule requiring some people to donate to charity. This is because the general acceptance of a code that includes such a rule would almost certainly result in better expected consequences than any code that lacked such a rule. How should this "rule regarding charity" function in the ideal code?

Imagine the following world:

World 1. In this world, the moral code includes a rule requiring everyone to give 10% of their annual income to charities that benefit the least well off in society.

The code in force in this world seems suboptimal, since it requires everyone, even those who are poor, to donate 10% of their income to charity. For the worse off in society, this would constitute a significant economic burden, and may end up making them *even* worse off. As a result, this rule is self-defeating and may end up making the world worse than if this rule were left out. This code cannot be the ideal code.

[3]I borrow this formulation from Kahn (2013).

The rule regarding charity will be self-defeating in this way only if our moral theory holds that the rich and poor are under identical obligations to give to charity. However, they are not. One way of avoiding this problem, then, is to craft a code of rules according to which our obligation to donate to charity is *conditional* on our being rich. Suppose that we move from World 1 to the following world:

> **World 2**. In this world, the moral code includes a *conditional* rule specifying an obligation to give to charity. The rule is to be phrased in roughly the following way, "*If you are at least moderately well off*, you ought to donate 10% of your income to charity."

The code in force in this world is morally better than the one in force in World 1 because its rule regarding charity requires agents to donate to charity only if they can afford to do so. As a result, it is not self-defeating, and we can expect World 2 to contain greater aggregate wellbeing than World 1. In fact, we can expect at least some in World 2 to be better off than they were in World 1 while no one is worse off.[4]

Including conditional rules in the ideal moral code is one way of capitalizing on the benefits of moral relativism without bearing the alleged theoretical costs. Some philosophers have pointed out that accepting moral relativism could do *even* more to promote wellbeing than teaching a single, universal code of rules, even one that includes conditional rules.[5] Imagine the following world:

> **World 3**. World 3 contains all the same agents as World 2. However, in World 3, society is to be divided along socioeconomic lines such that the at least moderately well off are to internalize one moral code—the "Rich Code"—and the poor are to internalize a different moral code—the "Poor Code." The Rich Code

[4]Hooker proposes this very solution at (2000: 87).

[5]Kahn (2012) gives two concrete examples of how this might be. His first example involves different populations spread across causally isolated possible worlds; his second involves causally isolated populations in the same world (2012: 7–10). I argue below that these conclusions likely hold even when considering agents who are not causally isolated. See also Kagan (2000), Mulgan (2005), Parfit (1984).

and the Poor Code differ in exactly one respect: the Rich Code requires its adherents to donate 10% of their income to charity; the Poor Code makes no demands regarding charity. As a result, Worlds 2 and 3 contain identical agents who are engaged in identical patterns of behavior and who have identical levels of wellbeing. (For the time being, I ask that we bracket the possibility of social mobility.)

If we are comfortable with moving from World 1 to World 2, then we must be comfortable with *relativized patterns of behavior* among agents, for in World 2, agents are effectively asked to act differently depending on their level of wealth. World 2 and World 3 are also identical in this regard, i.e. *from the outside*. Their differences become apparent only when we look more closely. There, we see that two individuals could be under the jurisdiction of two *distinct* moral codes, not just that they could variably satisfy the antecedent of some conditional rule. Yet it would be odd to suggest that for this reason World 3 is *morally worse* than World 2. By hypothesis, these worlds contain identical agents who are engaged in identical patterns of behavior and who have identical levels of wellbeing. In fact, it seems that World 3 is *not worse* than World 2.

Finally, imagine the following world:

World 4. World 4 is identical to World 3, except that in World 4, the Poor Code differs from the Rich Code in the following way: rather than simply removing a rule regarding charity from the Poor Code, we instead *substitute a new rule* in its place.

In World 3, the Poor Code differed from the Rich Code merely in lacking one of the Rich Code's rules. But removing a rule from the Poor Code leaves some psychological resources unused as the poor are required to internalize one less rule. In World 4, we capitalize on these unused psychological resources by including in the Poor Code a new rule. The move

from World 3 to World 4 promises to make at least some better off while making no one worse off. Here is how:

For rule-consequentialists, the expected value of everyone's internalizing a code is a function in part of the psychological cost of internalizing the code. This psychological cost in turn is a function of a code's size (i.e. number of rules), complexity, and distance from humans' natural tendencies (e.g. codes that are more demanding are more difficult to internalize). The expected consequences of everyone's internalizing a code is a function in part of the size and complexity of a code because a code that contained more fine-grained rules pertaining to the variety of moral situations we are likely to encounter would better equip an agent to maximize her contribution to wellbeing. Thus, it is almost always a suboptimal strategy for a code of rules to leave some psychological resources unused: aggregate welfare could almost always be increased by adding an additional rule to the ideal code.

The poor clearly have a reduced psychological burden in World 3 compared to World 2 because the Poor Code has one fewer rule while being otherwise identical to the Rich Code. The rich also plausibly have a reduced psychological burden because internalizing an unconditional rule is simpler than internalizing (and applying) a conditional rule. At any rate, the psychological burden on the rich is surely not higher in World 3 than it is in World 2. Hence, the total psychological cost is pareto reduced in World 3 compared to World 2. No one incurs a higher psychological cost and, because the same patterns of behavior that Hooker endorses in World 2 obtain in World 3, there is no change in the aggregate welfare in the world.

The Poor Code in World 3 leaves some psychological resources unused. We could make use of these newly freed up psychological resources by adding an additional rule to the Poor Code regarding, e.g., the treatment of other people or their property. If we are careful in designing this new addition to the Poor Code in World 4, its inculcation by part of society would plausibly make the world better off. Thus, fully embracing synchronic moral relativism

in World 4 more efficiently uses the available psychological resources in society and thereby makes the world better off for some while making it worse off for no one.

Thus, we have an argument for adopting *relative rule-consequentialism*, the theory in place in World 4:

1. World 2 is preferable to World 1.
2. World 3 is not worse than World 2.
3. World 4 is preferable to World 3.

$\therefore$ 4. World 4 is preferable to World 1.

According to relative rule-consequentialism,

(1) An action is wrong for an agent to perform if it is forbidden by the code of rules to which she is subject.

(2) An agent is subject to a code of rules if and only if her being subject to that code of rules is in accordance with the ideal arrangement of all codes of rules.

(3) An arrangement of codes of rules is the ideal arrangement of all codes of rules if and only if it is the arrangement that would have maximum expected consequences in terms of wellbeing.

3.2 Objections

Below I consider objections to premises (1) and (2). I do not anticipate objections to premise (3). Some brief preliminaries will be helpful before considering these objections. According to Hooker's formulation of rule-consequentialism,

> An act is wrong if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone everywhere in each new generation has maximum expected value in terms of wellbeing... (2000: 32)

Notice that Hooker's criterion of wrongness is consistent with the following position:

> *Diachronic moral relativism.* The moral status of an action depends in part on the time at which it is performed.

Hooker does not hold that the time at which actions are performed matters in itself. Rather, he holds that the moral status of an action may differ between generations because the code of rules to be internalized is to be reevaluated with each new generation.[6] But to hold *diachronic moral relativism*, we need only think that otherwise identical actions may be right if performed at one time and wrong if performed at another time.

Notice also from Hooker's locution "everyone everywhere" that he rejects[7] the following position:

> *Synchronic moral relativism.* An action performed at at time may differ with regard to its moral status depending on who performs it.

As we will see, Hooker's theory is inconsistent: he cannot countenance diachronic moral relativism while rejecting its synchronic sibling. Moreover, Hooker's arguments in favor of diachronic moral relativism often count in favor of synchronic moral relativism, exacerbating this inconsistency.

3.2.1 Objections to premise (1)

Consider the following objection: World 2 is not better than World 1. In World 1, the poor give to charities that benefit the least well off. But then they immediately receive this money back, since they themselves *are* the least well off. So, we could expect Worlds 1 and 2 to have identical levels of aggregate wellbeing.

It will be difficult to respond to this claim without waving my hand at some empirical assumptions. But since the objection itself is based on an empirical assumption, the question

[6] The reasons for this reevaluation are discussed below.
[7] The reasons he gives for this rejection are discussed below.

is ultimately whether we think its central assumption is more likely than what I will say now. First, it is not right to say that the least well off in society will 'immediately' have their money returned to them by the charity. Rather, more likely, they would receive this assistance in the next month, or the next quarter, or some such. In the interim, they would be needlessly impoverished. If these are people who would be made worse off by giving (away) their money to charity, then they are also people who will be made worse off by giving their money to charity for a few weeks or months. Second, it is unlikely that these people would receive the same amount of money that they donated in return. What is more likely is that they would receive slightly less than the original amount, once we account for the overhead of the charities.

3.2.2 Objections to premise (2)

Empirical objections to premise (2)

Several objections to premise (2) can be classed as *empirical* objections: they are objections that rely on claims about what World 2 and World 3 would really be like. Consider these objections:

1. It is unrealistic to expect World 2 and World 3 to contain identical levels of wellbeing. Hooker gives this argument at (2000: 82) where he says, "there are advantages of having just *one* code for internalization by everyone. These advantages include convenience" (italics in original). Hooker does not elaborate, but it seems the only way of understanding this is as a consequentialist consideration. Perhaps he means something like this: World 2 and World 3 would not have all the same agents engaged in identical patterns of behavior, since the moral code needs to be taught to different sub-groups of people. We can imagine how epistemically demanding this would be in practice: Imagine a kindergarten teacher or Sunday school teacher in World 3 separating her students by income level and asking half of them to cover their ears each time morality is brought up. Since this would not have to take place in World 2, these two worlds do not contain all the same agents engaged in *identical patterns*

of behavior. Since it would take some additional time and resources to teach different codes to different social groups, World 2 and World 3 would not contain identical agents with *identical levels of wellbeing*, either. In this case, it is *more convenient* for there to be a single universal moral code, rather than a regime of relativized codes.

2. World 2 and World 3 would not contain all the same agents with identical levels of wellbeing, since relativizing the moral code would reinforce a kind of caste system, and awareness of such a caste system would make people in society worse off.

3. The above description of World 3 is an oversimplification, since it brackets social mobility. Of course it is unrealistic to suppose that no one will ever move between social classes. When that happens, a regime of relativized moral codes would fail to guide a person's action in the right way.

I believe one response will be sufficient to answer all of these objections: Allow me to introduce a modified version of the argument that appeals not to multiple populations that are interacting, but rather appeals to multiple populations who are causally isolated in some robust way. I will suppose that they inhabit two different planets that are far from one another. We may imagine *divided* versions of the worlds given above, i.e. Divided World 1, Divided World 2, and so on, each of which contains two separate planets. Imagine that the inhabitants of these planets are indistinguishable from humans in terms of their rational and empathic capacities. Finally, imagine that these two populations of moral agents vary widely in their material circumstances. For this version of the argument, consider a rule regarding resource sharing (not just donations to charity) since the "poorer" of these two worlds is in a situation of extreme material scarcity.

Imagine in the first of these divided worlds that everyone is required to internalize a single moral code with an imperative rule regarding resource sharing. Imagine in the second world that this rule is made conditional, which would presumably improve the conditions of the beings on the poorer planet, perhaps by requiring more resources to be shared in common. Imagine, finally, that in Divided World 4, the populations of these two planets are

to internalize moral codes that are strikingly divergent, as would be best for each individual population. Consider how this response answers the objections above: First, there is no additional cost between Divided World 2 and Divided World 3 associated with teaching different moral codes to different social groups. Instead, entire planets are to internalize different moral codes. Second, if these two planets are causally isolated, then there will be no worries about the appearance of a caste system: the two populations simply do not know about each other. Third, and finally, since the two planets are causally isolated, there will be no mobility from one to the other, i.e. from one level of material scarcity to another.[8] Thus, there would be no worries about the possible consequences of social mobility.

This weakened version of the argument sacrifices something in terms of its provocativeness: it does not establish that, in worlds like ours, different social groups could be asked to internalize different moral codes. But it does still establish that various populations of moral agents, alive at the same time, could be required to internalize different, conflicting moral codes, and this would establish synchronic moral relativism all the same.

Theoretical objections to premise (2)

Many other potential objections to premise (2) could be classed as *theoretical* objections: these are objections about the theoretical reasons we have for preferring World 2 to World 3, rather than claims about what those worlds would be like in terms of their expected levels of aggregate wellbeing. I will respond to these objections in turn:

1. Rule-consequentialism is the best account of impartially justifiable morality.[9] But allowing two moral codes to be in place at once greatly undermines the ability of one person to justify her actions to another. In fact, we may even think that the *possibility* of impartial justification depends on there being a single moral code. Imagine two people under the

[8] If we suppose that faster than light travel is impossible, then any two planets orbiting different stars are *practically* causally isolated in the sense required for this example.

[9] I am indebted to Jussi Suikkanen for pointing out this objection.

jurisdiction of two moral codes who come to interact. Which moral code governs their interaction? At any rate, these two people will disagree about which action is right.

While this objection, I believe, is properly considered theoretical, it can be answered by the same response as the empirical objections above. In particular, the situation contemplated in this objection would not obtain in the divided worlds examples I have already discussed.

2. Hooker himself is not committed to the maximization of wellbeing, and so he has no theoretical reason for preferring World 3 to World 2. In fact, he has a strong theoretical commitment to moral universalism because this coheres well with our intuitions about the general structure of morality, and the idea that morality is a shared, collective project is one of rule-consequentialism's most attractive insights.

Let us take a closer look at why Hooker rejects moral relativism. He says:

> the idea of relativizing codes to groups is on the road to relativizing them to sub-groups, and at the end of that road is relativizing them to individuals. To go down that road is to turn our backs on one of the traditional attractions of rule-consequentialism—namely, its basis in the idea that morality should be thought of as a *collective, shared* code. (2000: 87)

Hooker approvingly quotes Bernard Gert in support of this claim:

> Morality should be taught to everyone, adherence to it should be endorsed by all members of society whose endorsement counts, and everyone should be urged to follow it. The requirements of morality apply to all rational persons. (1988: 216)

It's not clear that this quote provides the support that Hooker needs. First, the conception of morality as a system of universal rules is controversial among moral philosophers. The virtue ethics tradition challenges this conception. Maximizing act consequentialism might also be seen to challenge this conception. We might think of act consequentialism as

requiring adherence to a single moral rule: act so that you maximize your contribution to total expected wellbeing. However, maximizing act consequentialists also recognize it may be self-undermining for people to act with that rule in mind[10], and have accepted that not everyone ought be *taught* the same moral rules.[11] This raises an additional difficulty: it is unclear whether *morality* for Gert is a rigid or flaccid designator. Act consequentialists, even when they accept esoteric morality, might agree with Gert that 'morality' applies to all rational persons, but they might deny that it is *the same* morality in each case. No one gets a free pass from the demands of morality, but perhaps morality does not require of A the same thing it requires of B; perhaps A and B should not be taught the same system of rules; perhaps A and B ought to act with different rules in mind; and perhaps A and B ought to be judged according to different sets of rules. If this is all true, it is not clear what overlap is left for us to say that *the same* morality applies to both of them. In fact, Hooker accepts conditional rules, and so he already accepts that morality does not make identical demands on all its agents, even of those alive at the same time. Moreover, and most troubling, it seems insincere for Hooker to claim that morality is a system of rules that applies equally to all rational agents: According to what he says elsewhere, those in the previous and subsequent generations, though they be rational, need not follow the same rules as we! And this is true in the most robust sense: not merely that they are under the jurisdiction of some conditional rule, but that they are under the jurisdiction of an altogether different rule.

Hooker's worry is that taking the first step down the road to relativism leads inexorably to the following position:

> *Individual RC.* It is wrong for A to perform some action if that action is forbidden by the code of rules whose internalization by A would lead to maximum expected consequences in terms of wellbeing, impartially considered.[12]

[10]See Railton (1984).

[11]This is to adopt the view that morality may be *esoteric*, i.e. not publicized to all. See Norcross (1997: 387–391) and Eggleston (2013).

[12]Kahn calls this the *Strong Thesis*: "for each agent there is exactly one IMC [ideal moral code, which determines the right action *for that agent*]" Kahn (2012: 3).

If Individual RC collapsed into equivalence with some other moral theory, it would presumably inherit the vices (and virtues) of that other theory. Does Individual RC collapse into equivalence with any other moral theories? According to Individual RC, each person ought internalize the code of rules that, given her peculiar psychological makeup, will maximize her expected contribution to aggregate wellbeing, taking into account the code's internalization costs. Following Hooker, let's suppose that to internalize a rule involves forming the associated motivations to comply with the rule, but also involves

> *more than* certain associated motivations. It also involves having sensitivities, emotions, and beliefs—indeed a particular cast of character and conscience. If you *accept* a rule against stealing, you will be motivated not to steal simply because it is stealing... You will also be disposed to feel guilty if you steal, disposed to resent stealing by other people, and disposed to blame them for it. (2000: 91)

On its face, Individual RC bears a resemblance to some other forms of consequentialism: maximizing act consequentialism[13], sophisticated consequentialism, and motive utilitarianism. In fact, it does not collapse into content equivalence with maximizing act-consequentialism[14], sophisticated consequentialism, or motive utilitarianism.[15] Thus, Individual rule-consequentialism must be evaluated on its own merits.[16]

Consider one final theoretical objection to premise (2) from Hooker: The best argument for rule-consequentialism, he says at multiple places,[17] is that it does a better job than other theory of justifying and systematizing our most deeply held moral beliefs. This is the

[13]See Lyons (1965). Note that this is separate from the charge that rule-consequentialism as a *universal* theory collapses into extensional equivalence with maximizing act consequentialism. That is the standard "collapse objection," which can be found in Smart (1973) among others.

[14]Kahn ably demonstrates this at Kahn (2012: 12).

[15]These arguments are nuanced and overly technical. In order to avoid straying too far from the main point of the discussion here, I have moved these arguments to Appendix A.

[16]Consider, for example, another merit of rule-consequentialism that Individual RC would sacrifice. Since it would be more difficult, if Individual RC were publicized and generally followed, to confidently predict others' behavior, Individual RC would lack rule-consequentialism's advantage in solving collective action problems and the resulting beneficial "expectation effects."

[17]See Hooker (2000: 88, 101).

argument for rule-consequentialism from reflective equilibrium. These moral beliefs include beliefs about the general structure of morality, e.g. that morality is a universal, shared enterprise. Accepting moral relativism sacrifices coherence with this deeply held moral belief, and therefore weakens the justification for rule-consequentialism. Thus we pay a substantial theoretical price in moving from World 2 to World 3.

First, we might think it is an open question whether our commonsense morality is as opposed to moral relativism as Hooker would have us think. But this is not a strong response. After carefully considering their default relativistic position, many people come to reject it.

There are other worries about the theoretical costs we might pay in moving from World 2 to World 3. For example, we might think that World 3 is worse than World 2 according to Ockham's razor. If Worlds 2 and 3 have all the same agents who have identical levels of wellbeing, then perhaps we ought to prefer the world with fewer theoretical entities, i.e. fewer moral codes. Thus, World 2 would be preferable to World 3. The response I will offer now undermines all objections that trade on an alleged theoretical cost incurred in the move from World 2 to World 3.

Consider again the moves from World 2 to World 3 and from World 3 to World 4. In moving from World 2 to World 3, we relativize the ideal moral code to separate groups. In moving from World 3 to World 4, we introduce a new rule into one of these moral codes in order to increase expected wellbeing in World 4. Thus, the move from World 2 to World 3 can be understood as paying a theoretical price. The move from World 3 to World 4 can be understood as claiming a reward in terms of increased expected wellbeing. Thus it is easy to see why the move from World 2 to World 3 is especially objectionable for some: we pay a theoretical price with no payoff. That payoff only comes in the next world.

Let us consider, then, the move directly from World 2 to World 4. In making such a move, we would go directly from teaching the poor an otiose rule to teaching them a useful rule, without simply jettisoning the rule to begin with. In one fell swoop, then, we would pay a theoretical cost by relativizing in exchange for a moral benefit of making the world better.

Now we may weigh this cost and this benefit directly against one another. How strong is our commitment to moral universalism, really, such that we are satisfied to forego increased expected wellbeing? How stridently should we insist on keeping the world a worse place in order to maintain an abstract theoretical commitment? This is a question for another time, but I suspect there will be wide disagreement among the answers to this question.

In order to facilitate this comparison, we could even consider our reasons for moving from World 3 to World 4 as theoretical reasons rather than moral reasons. Here is how: In moving from World 3 to World 4, I suggested that we make some people better off while making no one worse off. Thus, World 4 is clearly morally better than World 3. Recall that the best argument for rule-consequentialism is the argument from reflective equilibrium, which argues that rule-consequentialism best coheres with our considered beliefs about morality. Presumably, many of us have a strong *theoretical* commitment to securing costless benefits. That is: we believe that if some theory can make some people better off while making no one worse off, this counts as a theoretical advantage of that theory.[18] Now, we can understand a direct move from World 2 to World 4, as a purely theoretical tradeoff: a tradeoff between our intuitions regarding the desirability of moral universalism and intuitions regarding the desirability of costless benefits. I suggest the case is at least a wash and may in fact favor relativizing the ideal code to social groups in order to secure such benefits. Thus, even an argument from reflective equilibrium that seeks to block the move to from World 2 to World 4 may end up supporting that very move.

Finally, notice that this same response can be generalized and given against any miscellaneous theoretical complaints about the move from World 2 to World 4: simply consider the alleged theoretical cost against the theoretical advantage of securing costless benefits in World 4.

[18]This theoretical advantage is *in addition* to making the world where everyone follows that theory *morally* better.

3.3 Hooker's support for diachronic moral relativism

We have seen that Hooker's embrace of diachronic moral relativism, given his rejection of synchronic moral relativism, is inconsistent. Let us consider the reasons Hooker gives for endorsing that view. First, the ideal code of rules might need to be changed so as to respond to new problems or to clear away rules meant to deal with 'old' problems that have been at least in some part *solved*, e.g. institutionalized slavery and the subjugation of women. In correspondence, Hooker has suggested global warming as a *new* problem that demands we alter the moral code we internalize and pass on. Presumably this is intended to maximize expected consequences in light of the state of the world and our knowledge about it. As we noted above, however, maximizing expected consequences could also be accomplished by relativizing the ideal code to social groups. If Hooker's rationale is that some generations will have to deal with problems that we don't currently have to deal with, we should recognize that this is also true of certain social groups, e.g., with regard to global warming.

Hooker's second argument concerns gradual changes in human psychology due to genetic engineering or genetic drift:

> Assume that new generations are not changed genetically. If genetic engineering alters human genetic makeup, the codes that are best will probably be different... (2000: 32 n1)

Here, Hooker is likely responding to possible changes in humans' capacities for empathy and other-directed actions, calculative capacities, introspective abilities, etc. Such changes, if significant, might alter the moral code that it would be best for everyone to internalize. But marked differences along these same lines exist in humans that are alive today. If changes in these factors militate in favor of changing the ideal code for future people, why not for some people who are alive now? If Hooker is sensitive to potential differences across human history in our psychological makeup, he ought to allow humans who presently differ psychologically

in important ways to internalize different codes of rules. And this just is synchronic moral relativism.

Hooker also supports recalculating the costs of internalizing the ideal code because of his desire to calculate such costs starting from *a blank slate* (2000: 79–80). In particular this is in order to ignore the costs of overcoming racist, sexist, or homophobic dispositions in such a calculation. Hooker gives two reasons in support of his position: (1) that doing so coheres with our common belief that racism, sexism, and homophobia are illegitimate attitudes unworthy of our respect (and so they are); and (2) that the satisfaction or frustration of 'external preferences' such as those regarding other people are not plausible constituents of a person's wellbeing, and so we need not consider them in our calculations of aggregate wellbeing. But Hooker need not commit to diachronic moral relativism to accomplish this: he might as well merely stipulate that overcoming these beliefs should not be included in calculating the internalization costs of codes of rules.

3.4 Conclusion

There does not seem to be a good reason for relativizing along generational lines and not at the same time relativizing along social lines. Hooker might answer that people who live in different generations are causally isolated from one another in a way that people who merely live in different societies (and are alive at the same time) are not.[19] But that is clearly false. The actions of the present generation can affect the lives of even distant generations. This is clearest in what Parfit calls "same person" cases, where our actions make some particular person who exists later better or worse off. But it is also arguably the case in "different person" or non-identity cases, where our actions can affect the quality of life of someone who comes to exist later even if she is not identical to the person who would have existed

[19]For the time being bracket worries about specifying where one generation ends and another begins. (On this note, the "same-generation-as," or at least the "alive-at-the-same-time-as" relation is not transitive.) In fact, society resembles a rope of overlapping chains of generations, such that we can never say when one generation ends and another begins.

otherwise (Parfit, 1984: 355–356). So it cannot be an appeal to causal separation that is doing the work here. Even supposing there were some kind of special causal separation between generations that did not exist between people alive at the same time: If we are worried that people alive at the same time who had internalized different sets of rules would causally interact, the only reason this could matter is if it made a difference to the expected consequences in that world. But it's not clear why we could not simply factor this into the calculus in determining which codes of rules to teach.

The only reason left for opposing synchronic moral relativism is dogma. Hooker is not alone in this regard, of course, but he makes himself uniquely vulnerable to this criticism by allowing into his theory a form of moral relativism along generational lines. Hooker uses a few bad words to describe the proposal for synchronic moral relativism, calling it "paternalistic duplicity" to suggest one code for the "enlightened elites" of society and another for the rabble (2000: 85). But why is this objectionable? We typically think of people who are *elitist* as selfish or arrogant; we resent economic elites for their avarice, we resent intellectual elites (insofar as professional philosophers are not they) for their sense of superiority. No wonder, then, that we would oppose any suggestion that *these people* should be allowed to internalize a code that is different, especially if we view that code as *more permissive*.

But this all ignores the opposite possibility. Why not consider the class of *moral elites* in a society as those people who are superior along dimensions of calculative abilities, empathy, and generosity? These are people who are more willing to sacrifice for the sake of others, and so a code relativized to their psychological nature would actually be *more demanding*. Hence, allowing them to internalize an elitist code such as this would be better for everyone. Moreover, by putting these moral elites on display, it might set an example for the rest of society. Even though these moral elites are still playing by a different set of rules in this scenario, our worries about paternalism and duplicity are greatly diminished.

I have argued that rule-consequentialists with certain theoretical motivations ought to adopt a form of moral relativism, according to which multiple moral codes are in place at once. This is true for theorists who support rule-consequentialism for consequentialist reasons but, as I suggested above, also applies to theorists who hold rule-consequentialism because of how it fares in a test of reflective equilibrium. We ought to weigh our theoretical commitments to universalism—which are admittedly usually very strong—against our commitments to making people better off at no one's expense. I singled out Hooker as one theorist who avoids synchronic moral relativism at a significant cost to wellbeing. But he has already let in its theoretical sibling, diachronic moral relativism. In light of the arguments given here, I suggest it may be best for him to abandon his commitment to diachronic moral relativism, which could be done with only a slight change to his criterion of rightness. If he is unwilling to jettison the one, then he should accept both simultaneously.

Chapter 4

A *jus in bello* rule-consequentialist code of morality

The collection of moral constraints governing warfare can seem like a hodgepodge of unrelated rules: soldiers are required to wear a fixed sign visible at a distance, required to give quarter to surrendering troops, and required to minimize civilian casualties, among other things. Is there some single principle that underlies these and the other rules of warfare and, in doing so, lends them both their content and justification?

We stand to gain two things by identifying such an underlying rationale. First, any moral theory that can unify its disparate commands under a single rationale is a more elegant, and therefore more desirable, theory. Second, we could use such a theory to provide a general—yet informative—theory of the moral psychology and decision procedure of the just warrior, thus strengthening the ability of soldiers and policymakers to make principled decisions where the rules of warfare do not clearly settle the case.

I suggest that rule-consequentialism is especially well-suited to this project of unifying and justifying the rules of war. This is because rule-consequentialism shares a remarkable structural similarity with a plausible theory of *in bello* morality, namely, as a set of near-absolute rules chosen with reference to and justified by some consequentialist goal. For

warfare, I suggest that goal is *minimizing the horror* of war. Accordingly, I sketch a rule-consequentialist *in bello* code of morality. I will also discuss the moral dispositions of a soldier who has successfully internalized this moral code, i.e. the *conscience* she would have. Finally, I will discuss the implications of this view for three contentious topics in the military ethics tradition: the doctrine of double effect, supreme emergency, and the problem of noncompliance. We will see that this theory offers a plausible justification of the doctrine of double effect. We will also see that rule-consequentialism already boasts the conceptual resources to bring clarity to the notion of supreme emergency, namely, in the form of an "avoid disasters" clause that triggers in the face of especially catastrophic threats.[1] Lastly, I argue that Walzer's principle of supreme emergency is too restrictive and the typical rule-consequentialist view too permissive when faced with noncompliance, and suggest taking a middle ground. The result of all of this is the beginning of an original, nuanced, and plausible unified view of *in bello* morality.

4.1 Why look for a unified theory of military ethics?

Soldiers and statesman are bound by a multitude of moral constraints in warfare. These constraints can seem like a hodgepodge, or to borrow a phrase from elsewhere in normative ethics, like an 'unconnected heap of duties' (Hooker, 1996). Unifying this unconnected heap could benefit us in at least two ways.

First, we would gain at the theoretical level. Suppose we are considering two theories, A and B, which issue equally informative and intuitively plausible moral commands. So far, A and B are equally appealing candidate moral theories. Suppose, however, that theory A also boasts an underlying, unifying rationale for its commands while B lacks any such rationale. Proponents of theory A can point to reasons why its commands are commands, "all the way to the ground," as it were. Theory A would have an advantage over theory B in this case;

[1] In Appendix B, I give further reasons for preferring a consequentialist axiology to a communitarian one when formulating a theory of supreme emergency.

it could claim a theoretical virtue that theory B could not claim. Other things being equal, then, theory A would be a better theory, and we would be right to prefer it over theory B. I suggest that our current theories of *jus in bello* are in the position of theory B. If a *jus in bello* code of morality could be derived from a single, unifying moral principle that was independently plausible, then our theories of the ethics of war would be improved.

Perhaps some of the above has gone by too quickly. Aren't there already very explicit underlying goals of the just war tradition? Clearly, there are some rules of warfare that hail from some common origin, and we have a common goal in erecting and enforcing these rules as constraints. Namely, there exist the general goals of proportionality and discrimination in the conduct of warfare, which in turn generate more fine-grained, concrete moral requirements. These underlying goals do help to explain the constraints at the surface level of *jus in bello*, but they do not explain all of the surface constraints. Moreover, the requirements of discrimination and proportionality themselves are not basic, and admit of further unification. Perhaps both can be explained by an appeal to the value of human life. But the just war tradition does not hold that all human life is equally valuable, as evidenced by the special protection afforded to noncombatants.

If we could find a unifying, independently plausible justification that grounded the moral constraints on warfare, our theory would be more appealing in itself. But theoretical concerns like this are controversial, and some would not treat it as even a tie breaking reason for preferring one theory over another. It must be admitted, however, that deriving such an underlying rationale would yield dividends at the practical level by helping us divine the answers to difficult or controversial questions. Military ethics is rife with such cases, and identifying an underlying rationale could shed light on these. It could provide guidance in cases such as supreme emergency, the doctrine of double effect, and noncompliance—all three of which are discussed below. It could also guide us when evaluating emerging military technologies, such as cyberweapons or autonomous weapons systems, which have often stymied consensus among military ethicists.

4.2 A unifying rationale

So, there are good reasons for seeking an underlying, unifying rationale to our theories. When discussing an underlying, unifying rationale, I have in mind something like Kant's categorical imperative. While Kant gave several formulations of his categorical imperative, they were allegedly interderivable.[2] Let's take Kant's first formulation: "Act only on that maxim by which you could at the same time consistently will that it become a universal law of nature." If Kant is to be believed, this single principle itself is a complete picture of morality; it tells us everything we need to know. It tells us how to act in each case and it goes some way to justifying those actions. While some additional details are required to justify the principle all the way to the ground—such as the preeminent role *rationality* plays in Kant's moral theory—we have a single principle that lends everyday moral requirements both their content and their justification. It tells us both *what* to do and *why* to do it. What we seek is an analogous principle to unify and justify the rules of warfare.

4.2.1 Just war theory as a system of near-absolute constraints on warfare

First, the rules of war are a set of near-absolute constraints. For example, the prohibitions against intentionally targeting noncombatants, taking human shields, and mistreating prisoners of war are extremely stringent. In fact, there are very good reasons for soldiers and policymakers to believe that the rules of war are *actually* absolute: problems would likely arise if we were to see ourselves as free to break them. One problem of particular concern is the increased threat of escalating, violent reprisals that would result. Motivated by this fear, Walzer and many other just war theorists therefore hold that we are never at liberty to violate the rules of war.

Even given this description of the theoretical landscape, where these rules are often *described* as absolute, they are not absolute in practice. There is hardly a more egregious vi-

[2]While this seems relatively plausible for some pairs of formulations, such as the first and second, contemporary Kantians are still unsuccessful in deriving *every* one of them from *every* other.

olation of the rules of warfare that intentionally murdering large numbers of noncombatants, for example, by intentionally bombing city centers. Thus, Walzer, among others, considers the prohibition against city bombing absolute. Even though Walzer is infamous for his defense of *supreme emergencies*, cases in which we are permitted to violate the rules of warfare, he still believes in these cases that the rules of war do not recede and are not overridden, but are *violated*.[3] Thus even in the hands of someone like Walzer, who takes the rules of warfare seriously enough to deny that they genuinely recede in supreme emergencies, we find the practical injunction to violate those rules. Thus the rules of war are only near-absolute.

4.2.2 The deontologist's challenge

If we wish to generate and justify the rules of war, we will need to settle on a theory whose structure is as close as possible to just war theory, i.e. we would do best to find a theory that is also a set of near-absolute constraints. Ideally we would find a model elsewhere in normative ethics that can supply us with a theoretical framework, a theory that is amenable to what we have said in the previous section about the structure of just war theory. It might seem easy to narrow our focus immediately to some species of deontology. In fact, if the choice is between consequentialism and deontology, the deontologist might feel she has a leg up on the consequentialist. She might say:

> This is an easy win for the deontologist. Not only have deontologists included constraints in their theories—and often very strict constraints—but the idea of constraints is *antithetical* to the consequentialist project! We do not have to look far at all to find just war theorists rejecting consequentialism nearly out of hand, motivated by this belief. Thus, if we are to narrow our focus to some corner of normative ethics, we should limit ourselves to considering various species of deontology.

[3]Walzer's description of supreme emergency is puzzling in part because of this explicitly paradoxical feature: we at least occasionally *ought (all things considered)* to act in some ways that will knowingly violate the rules of warfare.

This is the *deontologist's challenge* and, if it were accurate, it would quickly rule out consequentialist theories from our consideration. The deontologist has good reasons to believe in a profound tension between consequentialism and just war theory, a tension which has not been lost on others. But the challenge is only successful if it is true that, as the deontologist says, constraints are antithetical to the consequentialist project.

The central complaint with consequentialism from just war theorists is an ignorance of the differential value of the lives of combatants and noncombatants, an ignorance of the unanimous view among just war theorists that, in wartime, we cannot trade off the lives of one soldier against one combatant. Nor is it the case—some[4] have pointed out—that we can trade off the lives of a thousand, or even a million soldiers against the life of one noncombatant. Noncombatants are *off limits* from intentional attack, no matter the value or strategic advantage that may be secured in intentionally targeting them. Consequentialism allegedly can make no sense of this, since all are to count for one and no one for more than one. For these reasons, consequentialism has been considered at least misguided and at worst an enemy of the just war tradition, ready at a moment's notice to license wholesale slaughter or, in Nagel's terms, *massacre*, in the interest of mere expediency.

Careful students of consequentialism will notice that one species of consequentialist theories has been neglected in the above discussion. The characterization of consequentialism that our deontologist gives is not true of every species of consequentialism, but only those species that are *direct*.[5] For while most consequentialists reject constraints, not all of them do. Among the genus of indirect consequentialist theories that the deontologist hastily rules out is one notable species: rule-consequentialism. Rule-consequentialism (RC) is a family of ethical views that are committed to the following two principles[6]:

[4]Anscombe (1981).

[5]And even then it's not obvious that all she says is correct. For example, see the work of Doug Portmore, which has focused on developing an axiology that could allow consequentialists to replicate the requirements and prohibitions of deontology Portmore (2007). See McNaughton and Rawling (1991; 1998) for opposing discussion of the same question. Work by Portmore and others has led to the belief, now fairly common, that any deontologist theory may be "consequentialized," that is, reproduced as a species of consequentialism.

[6]I paraphrase this formulation from Kahn (2013).

(1) An action is right if and only if it accords with an ideal code of moral rules.

(2) A code of rules is ideal only if its acceptance[7] by some portion of the population (usually at least the overwhelming majority of agents) would have the best expected consequences.

In contemporary forms of rule-consequentialism, including the version espoused by Hooker (2000), the theory's foremost contemporary defender, the ideal moral code will contain many near-absolute moral rules. Rules against harming others or their property, including murder and theft, are plausible candidates for rules that would be near-absolute. This is because we could expect a code of rules that prohibited those actions, except in very rare circumstances, to make the world better, were it internalized by most agents. Thus indirect consequentialist theories like rule-consequentialism maintain a place for near-absolute constraints.[8] Rule-consequentialism thus avoids the deontologist's challenge.

4.2.3 The ultimate goal of the just war tradition

The deontologist may respond:

> Rule-consequentialism does let in near-absolute constraints, but it is still a non-starter as a model for the rules of war because it relies on there being some consequentialist criterion by which we judge a set of rules to be the best. Thus we can only understand just war theory as a species of rule-consequentialism if we can specify the consequentialist criterion by which we judge candidate codes governing the conduct of war. To be fully convincing, this consequentialist criterion must be independently plausible as something worth pursuing, and it should bear a significant resemblance to the project of just war theory. There exists no such

[7] Distinctions between internalizing, accepting, and following the ideal moral code are real but need not concern us here.

[8] This could be true of other forms of indirect consequentialism, for example, Adams' motive utilitarianism (1976). However, no other indirect consequentialist theory as clearly and explicitly consists of a set of near-absolute rules.

general consequentialist project to be found underlying the just war tradition. Thus, rule-consequentialism must be rejected.

The deontologist is correct that the work for rule-consequentialists is not yet done, but this second challenge is not difficult to meet. Just war theory can in fact be re-described as pursuing a consequentialist project, both in its *ad bellum* and *in bello* requirements. This consequentialist criterion is both independently plausible and bears a significant resemblance to the project of just war theory. I suggest that this consequentialist criterion is *the minimization of the horror of warfare*.

> *Horror*. Instances of suffering and death that are especially gruesome or tragic, beyond what is typical in human life.

Not all instances of suffering are horrible, but all things that are horrible in this sense are instances of suffering. It is meant to conjure that revulsion and visceral shock that is the natural response to reports of casualties in wartime: soldiers who are mangled, burned, disemboweled; innocents who are crushed or obliterated in their homes by night time bombing raids.

This definition is intuitively plausible. Consider these synonyms for *horrible* offered by Google, many of them equally apt descriptions of the phenomenon I wish to capture:

> dreadful, awful, terrible, shocking, appalling... horrendous, horrid, hideous, grisly, ghastly, gruesome, gory, harrowing, heinous, vile, unspeakable; nightmarish, macabre, spine-chilling, blood-curdling; loathsome, monstrous, abhorrent, hateful, hellish, execrable, abominable, atrocious, sickening, foul...

Reducing the amount of horror, so understood, is a worthy goal on its own.[9] It is a project we can agree we ought to pursue, whether we are consequentialists or deontologists.

[9]The reason we ought to reduce the amount of *horror* in the world might be parasitic on the reason we have to reduce the amount of *suffering* in the world. At any rate, we have good reason to reduce the horror in the world, whether this moral requirement is basic.

And many of the traditional constraints of just war theory can be understood as trying to minimize the horror of warfare. For example, take several of just war theory's traditional *ad bellum* criteria: last resort, probability of success, and proportionality—these all plausibly have consequentialist underpinnings. Since war is horrible, it ought to be a last resort, only pursued after all other (less-horrible) means of conflict resolution have been exhausted. Likewise for reasonable probability of success—since it would be especially horrible for a war to be waged in vain. (Notwithstanding that some theorists allow exceptions for heroic self-defense, even when foreseeably futile.) This is also the case with proper authority, though the case to be made is more circuitous here. While I am confident that a robust and plausible *ad bellum* morality premised on minimizing the horror of war could be developed, My interest here is mainly with developing an *in bello* code of morality, so I will have to leave this discussion unfinished.

Finally, here is a sketch of a rule-consequentialist theory of *jus in bello*:

> Rule-consequentialist *In Bello* Code. An act in wartime is wrong if and only if it is forbidden by the code of rules whose internalization by soldiers and policymakers we could expect to best minimize the horror of war.

The above criterion of wrongness can plausibly be used to underpin a regime of near-absolute constraints on soldiers and policymakers. These constraints are broadly consonant with the general principles of *jus in bello*: First, acts of force must *discriminate* between combatants and noncombatants. Extreme care must be taken to avoid injuries to noncombatants since the harm that befalls noncombatants epitomizes the horror of war. It may strike some readers as incredible that a consequentialist theory could justify an absolute moral distinction between combatants and noncombatants. Surely, a cornerstone of consequentialism from Bentham on has been to regard all individuals as morally equal and, more precisely, to claim that an episode of suffering is not more or less morally important depending on *whose* it is. But this objection mistakes the difference between mere suffering and *horror*, the more complex concept that tracks considerations besides mere suffering. Above

I suggested that *horror* tracks the degree to which harms are caused indiscriminately by warfare. Thus someone who has internalized the *in bello* code will strongly prefer the death of a combatant to a noncombatant, other things being equal—and even when other things are quite unequal.

Second, acts of force must be *proportionate* to the strategic goal, never inflicting more harm than is necessary, as one of the aspects of war that is horrible is its seeming disproportionality. See, for example, the absurd casualties endured in the trenches during the First World War, where thousands of soldiers could die in order to advance their state's territory by just a few miles, or less. Both of these constraints would also serve to minimize reprisals and escalations as far as possible, again reducing the total amount of carnage that takes place in wartime. This *in bello* code would also likely require the humane treatment of surrendering soldiers and prisoners of war, and respecting the other *jus in bello* constraints.

Notice that these near-absolute constraints can be understood as the corollaries of near-absolute rights, possessed both by soldiers fighting a war and by noncombatants who are too often the victims of war's destructive potential. We may call these quasi-rights, since they are justified only by the fact that their general observance would tend to minimize the horror of war, and they are not grounded in any of the bases more popular with deontologists, such as human dignity, rationality, or autonomy. However, this difference does nothing to diminish their stringency. Thus, this rule-consequentialist *in bello* code can effectively impersonate a just war theory grounded in a claim about the rights of soldiers and noncombatants. That is, this theory would plausibly be extensionally equivalent to a rights-based just war theory, and its analogous commands would boast equivalent strength.

For rule-consequentialism, the ideal moral code is the one that we could expect to make the world as good as possible if it were internalized by the overwhelming majority of the relevant moral community. For previous versions of rule-consequentialism in the normative ethics literature, this is usually taken to mean all moral agents. For an *in bello* rule-consequentialist code of morality, the relevant moral community is likely those people directly involved in the

business of making war, at least those in whom some decision making authority rests. As the criterion of wrongness above lets on, soldiers and policymakers are plausible candidates for the population who should internalize these rules. There are others who are nonetheless involved in the business of making war, such as those working in bomb factories, but they need not internalize the *in bello* code. This is for a few reasons: First, their decisions do not significantly impact the course of a war, at least not in the way that the decisions of soldiers and policymakers do. These people do not directly and immediately contribute to the suffering and death that war brings. And perhaps most importantly, their decisions and contributions to the war-making effort do not have implications for the army's respect for discrimination and proportionality. A factory worker surely cannot do anything herself to ensure that her bomb falls on—or is at least intended for—a combatant rather than a noncombatant. Thus, the population meant to internalize this code must be those actively engaged in the immediate business of waging war and those others whose decisions in large part determine the actions of the first.[10]

According to contemporary versions of rule-consequentialism,[11] to *internalize* a rule is to acquire a disposition to follow that rule, to feel bad when one breaks the rule, and to enforce general rule-following with social sanctions. The strength of the disposition to follow a rule is proportional to the importance of the rule itself, such that we ought to have a stronger aversion to murdering than to lying, for example. Likewise for the *in bello* code, to internalize a rule is to form the dispositions to avoid specific violations. This includes having the strong motivation to follow that rule for its own sake, to encourage others to follow it, to feel guilty when having broken it, and to censure others who have broken it. Since almost

[10]It was suggested to me that some factory workers could purposefully sabotage their work in the interest of weakening the war machine, and thus could have some impact on the justice of the war being fought. This suggestion is intriguing but I ultimately find it implausible. Briefly, I find it implausible that this kind of behavior could go on very long, carried out by enough people, in order to make a difference to the efficacy of the host country's war machine. It seems likely to me that such an occurrence would not go long unnoticed, and that it could very well result in the factory overseers' demanding even more work from their workers as punishment or in order to compensate for the erstwhile shortfall.

[11]Here I am thinking especially of Hooker (2000).

all of the *in bello* constraints are nearly absolute, soldiers who have successfully internalized the *in bello* code will find it nearly impossible to violate them.

In many scenarios in warfare, the received moral guidelines do not clearly settle the case for or against some behavior. Here, an *in bello* code offers a useful decision procedure for determining what an agent ought to do. Suppose that an agent is deciding between actions A and B. According to rule-consequentialism, she ought to ask herself, "Would the world be better if, in this situation, people felt compelled to do A or B?" Likewise, soldiers and policymakers who have internalized the *in bello* code ought to ask, "Would war be less horrific if, in this situation, soldiers and policymakers felt compelled to do A or B?" We should try to inculcate in our soldiers and policymakers those consciences such that, if all other soldiers and policymakers had the same conscience, war would go *best* in terms of being *least horrific*.

Finally, notice that the *in bello* code specifies the necessary and sufficient criterion of wrongness for acts of force during wartime. It seems that all actions that are wrong will be actions that violate some stringent right of non-combatants or otherwise run afoul of the *in bello* requirement of proportionality. And all actions that violate the immunity of non-combatants or are purposefully disproportionately violent are, of course, wrong in war. What we have now are the beginnings of a plausible unified account of right conduct in wartime.

4.3 The *in bello* code and supreme emergency

Let us now apply this fledgling theory to some of the more contentious topics in the ethics of war. Examining the theory's dictates here will provide the first evidence of its acceptability, or else cause us to reject it.

First, we will examine this theory's implications for supreme emergency. While Churchill is credited with coining the term *supreme emergency*, the doctrine receives its original philosophical explication in Walzer (1977). There, Walzer contemplates whether a potential outcome in warfare can ever be so cataclysmic as to justify massive and intentional viola-

tions of the rules of war. He believes that there are such scenarios, and the possibility of a Nazi conquest of Europe is his motivating example. Nazism, Walzer says,

> was an ultimate threat to everything decent in our lives, an ideology and a practice of domination so murderous, so degrading even to those who might survive, that the consequences of its final victory were literally beyond calculation, immeasurably awful. We see it—and I don't use the phrase lightly—as evil objectified in the world, and in a form so potent and apparent that there could never have been anything to do but fight against it... Here was a threat to human values so radical that its imminence would surely constitute a supreme emergency... (Walzer, 1977: 253)

Nazism's triumph would have represented the eradication of multiple political communities and would have threatened political liberalism the world over. In the face of such an imminent disaster, Walzer argues we are justified in violating the inviolable immunity of noncombatants; we are justified in doing whatever is necessary to avoid such a disaster. The normal rules of warfare do not recede or lose their force according to Walzer. On the contrary, Walzer relishes the apparent paradox in the idea that we could be *required to do what is morally wrong.*

Notably, rule-consequentialism already contains the theoretical machinery analogous to Walzer's supreme emergency. In rule-consequentialism, this exception clause takes the form of a rule stating simply, "Avoid disasters." The justification for the inclusion of this rule is identical to the justification for the inclusion of all of the other rules. Namely, everyone's internalizing a code that included a rule to avoid disasters would make the world go better than if everyone internalized a competing moral code that lacked such a rule, other things being equal (Hooker, 2000: 134–136).

Hooker's discussion of what constitutes a disaster in rule-consequentialism is intentionally vague. This is one place where vagueness in our moral theories may be inescapable, and where the best we can do is appeal to the judgment of a morally well-educated person.

He does tell us that greater disasters license the violation of more stringent rules. For example, presumably, we are justified in telling a lie to prevent a person from being harmed, and justified in harming one person to save the life of another. But the violations we are contemplating here are much more awful: we are contemplating intentionally killing and maiming not just a single person, but *scores* of *innocent* people, perhaps even hundreds of thousands of them, as was the case in the struggle against Nazism.[12] What scale of disaster could possibly justify this much wanton violence?

We will have to turn to Walzer for some insight into the conditions that justify resort to supreme emergency. Here is what Walzer (2004: 50) and others say. First, the threatened disaster must represent a "far greater immorality" than the actions we will undertake. In the case of Nazi Germany, this condition was met at one time since the total destruction or enslavement of a political community is much worse than the deliberate killing of noncombatants, short of the destruction of a political community. Walzer says that "our deepest values and our collective survival" must be at risk—presumably the "we" is relativized to a particular political community, i.e., Britain (Walzer, 2004: 33). Orend uses a less demanding criterion here, not invoking the imminent destruction of a political community, but instead using the terms, "widespread massacre and enslavement" (Orend, 2006: 140). Orend's position is weaker than Walzer's, and thus to justify a supreme emergency, Orend must rely on the more dubious claim that *widespread massacre and enslavement* is a far greater immorality than the *widespread deliberate killing of noncombatants*.

Second, the threat must be very real and very near success. Orend says that there must exist "public proof that the aggressor is just about to defeat the victim militarily" (Orend, 2006: 140). Later, Orend strengthens this requirement to say that it demands imminent military collapse, not just defeat. This, he means, amounts to the "total collapse of an effective armed forces capability, literally rendering [the victim] defenseless" (2006: 147).

[12]To be precise, Walzer believes that most of the bombings of Nazi cities were unjustified, since they took place long after they would have been necessary to repel a Nazi invasion of England (Walzer, 1977: 253 ff.).

Presumably this is abhorrent because it would pave the way for the widespread human rights violations contemplated in the first criterion.

Third, the immoral response being considered must be the only way of preventing the disaster from obtaining. Orend suspects that, regarding this last point, political leaders are much too quick to resort to supreme emergency because of a lack of imagination (2006: 156–157). Notice that this requirement is likely vulnerable to the same objections as traditional just war theory's "last resort" criterion. That is, the notion of *last resort* may be incoherent in the same way as the notion of a largest number.

In the hands of Walzer, the supreme emergency exception is only justified by a communitarian axiology. No simple conquest or loss of territory can justify a supreme emergency. The threat must be to the "ongoingness" of a community itself. Walzer's motivation for supreme emergency is to avoid a disaster on the scale of the destruction of an entire community and an ultimate loss of value. Threatened with eradication, Walzer claims that governments are absolutely forbidden from forsaking their citizens, "putting the community itself and all its members at risk," while there may be some options still open for resisting. Orend agrees, saying that a government standing down in the face of a supreme emergency is an abdication of its duty to protect its subjects "when effective resistance might still be available" (2006: 152). Walzer admits this argument fails unless we appreciate the irreducible value of the ongoingness of a political community. He explains:

> We do try to carry on, and also to improve upon, a way of life handed down by our ancestors, and we do hope for recognizable descendants, carrying on and improving upon our own way of life. This commitment to continuity across generations is a very powerful feature of human life, and it is embodied in the community. When our community is threatened, not just in its present territorial extension or governmental structure or prestige or honor, but in what we might think of as its ongoingness, then we face a loss that is greater than any we can imagine. (Walzer, 2004: 43)

Walzer refuses to say his argument depends on the possibility that, in the case of a political community, "the whole is greater than the sum of its parts" (2004: 42). However, Walzer must locate a special kind of value somewhere. Imagine two scenarios, one that involves the massacre of 200,000 noncombatants and a second that involves the massacre of 200,000 noncombatants who together fully constitute a political community. Supposing the other criteria for supreme emergency are met, Walzer's arguments above would support resort to supreme emergency only in the second case, even though these two cases contain equal amounts of carnage. What makes the difference is the threat to the political community itself in the second case. Walzer denies that his view places value in a political community itself, but it is difficult to see how he can avoid this. Walzer says that resort to supreme emergency is not permissible "when anything less than the ongoingness of the community is at stake" (2004: 46). At any rate, Walzer must locate value in the political community that is not simply the sum of the value of its individual constituents. That value is located in the ongoingness, the preservation, of a particular way of life.

Walzer's communitarian language is foreign to the consequentialist tradition, and his axiology is one it would be difficult to integrate into a consequentialist framework. Thus, a rule-consequentialist account of supreme emergency will differ from Walzer's. Luckily, there is ample independent reason to jettison Walzer's communitarian approach to justifying supreme emergency. First, we might think that tying supreme emergency to the death of a political community constrains the range of ethical judgments we can make during wartime. It would be troubling if supreme emergencies arose in all and only those situations in which the ongoingness of a political community were at stake. That's because it would prevent us from saying the following things. For one, certain political communities ought *not* be preserved, or at least the value of their preservation is less than that of some other political community. This plausibly depends on the size of a political community and the degree to which it either respects the rights or promotes the flourishing of its citizens. These are claims that Walzer cannot account for, but that a consequentialist theory of supreme emergency

could easily handle.[13] Conversely, it seems that a resort to supreme emergency might be justified by something short of the death of a political community. For example, Orend notes that the detonation of several atomic weapons in America might be enough to cripple its military responsiveness and damage its culture and way of life (2006: 145). While this would be an unimaginable catastrophe, it would not obviously spell the end of America's political community. Thus there seem to be disasters short of the destruction of a political community that could justify a resort to supreme emergency.

By jettisoning a commitment to the value of communities as such, we could also avoid many thorny ontological questions: What constitutes a political community and what are the conditions for its identity over time? What counts as threatening the destruction of a community rather than merely changing it significantly? Is slaughter and enslavement always required, as Orend says, or could a regime of draconian punishments (short of massacre) aimed at changing a society's mores constitute a sufficient threat to a culture? I criticize one popular argument for the value of communities as such in Appendix B.

Second, we could expand our theory of supreme emergency to apply to soldiers as well as policymakers. Walzer holds that supreme emergency is a doctrine only for political leaders, and that soldiers are never to take these kinds of matters into their own hands. But there may be situations in which a soldier has to make a decision to torture or to kill innocents to prevent wholesale rights violations, such as to prevent the threatened detonation of a nuclear weapon in a population center. In cases in which the soldier alone is faced with such a decision, and has lost contact with their superiors, it seems unwise to tie the hands of the soldier.[14] Notably, Orend is either sloppy on this point, or disagrees with Walzer without

[13]Walzer and Orend voice the standard position within the just war community that communities that are currently involved in massive human rights violations are not entitled to exist. But a consequentialist view could improve upon this sharp dichotomy between worthy and unworthy regimes, more closely matching the intuitive view that political communities fall along a continuum, and even those that are not currently involved in massive violations of human rights might be reprehensible enough that they are not due the same respect as maximally liberal democracies.

[14]This is not to give a blank check to soldiers to violate the rules of warfare. Recall that, ideally, the rules of warfare are near-absolute, and should be violated only in the most extreme circumstances, and only when the agent has excellent evidence of her justification.

being explicit: he says several times (e.g. 2006: 154) that supreme emergency is a doctrine for *soldiers* and policymakers, apparently without noticing his important disagreement with Walzer.

In fact, Walzer's strategy has been rejected by others, including Orend. Orend is notably reluctant to use the same kind of communitarian language that Walzer is fond of. Nowhere in his discussion of supreme emergency (2006: ch. 5) does he speak of the death of the community, or the end of its ongoingness. I take Orend's implicit rejection of Walzer to be a result of his recognition that such a communitarian approach is vulnerable to serious objections. A consequentialist view can make sense of the same three criteria just discussed for supreme emergency, and a rule-consequentialist theory of *in bello* morality can justify violating the otherwise-absolute rules of warfare. So much, then, for supreme emergency.

4.4 The *in bello* code and the doctrine of double effect

Once we understand a rule-consequentialist system as supplying the ideal shape of an agent's conscience, we can make sense of the importance of the mental states of soldiers and policymaker. This allows us to discuss in detail the doctrine of double effect (DDE), a notoriously contentious issue within just war theory. The doctrine of double effect can be taken either as a metaphysical thesis about the nature of harm or as a question of whether the mere mental state of an actor can affect the moral status of her action. The question before us here is one of the second type: whether the *in bello* code would include a stronger prohibition against *intending harm* to civilians as a means to achieving some end, or whether soldiers ought to feel *equally reluctant* to kill civilians when it is intended as when it is merely foreseen as a consequence of their action.

For consequentialists, many moral questions are simply empirical questions, in the sense that they could not be definitively answered from the armchair. The DDE is one such question. There are plausible consequentialist reasons on either side of supporting a stronger prohibition on intending harm than on merely foreseeing harm. First, we might think that,

because humans are imperfect rational agents, we should err on the side of caution in the case of intending harm. Both Walzer (2004: 38–39) and **?**: 112 ff, in fact, have argued in this way in analogous discussions of rule-violations.

On the other hand, there are many reasons to think that including an equal prohibition against intending harm as against foreseeing harm would better minimize the horror of war. For one, distinguishing between the two might be superfluous to the moral code. The existing demand that violent actions should be proportionate to their ends will *already* require that the number of noncombatant deaths is as low as is possible. We cannot be sure that a prohibition requiring soldiers to take additional care would be effective in significantly reducing the number of lives saved. In fact, we might think soldiers are better calculators when this stronger prohibition is not in place. Perhaps the willingness to intend harm is less likely to be abused because soldiers are more likely to take the deaths they inflict seriously when they are intended rather than when they are 'merely' foreseen. This is a reason we should allow them to contemplate intentional killings.[15]

The answer to this question ultimately relies on unavailable empirical evidence. How frequently does the opportunity to intend harm, versus merely foresee it, arise in war? How much of a difference would such a prohibition actually make? And so on. The stance we take on the doctrine of double effect from the point of view of *in bello* code depends on these questions, but their answers are impossible to discover from the armchair. For what it's worth, I find it plausible that, in the absence of empirical evidence to settle the claim, the best strategy is to forbid soldiers and policymakers from *intending harm as a means* for fear that might increase the horror of war. Thus, this theory might be termed *weakly* pro-DDE.

[15]In fact, if this is true, then we might take the argument to the extreme and say that there ought to be a stronger prohibition against merely foreseen civilian deaths versus than those that are intended. This would be because we want to guard against solders' writing off civilian deaths as an allowable level of 'collateral damage' when those deaths are merely foreseen.

4.5 The *in bello* code and non-compliance

Finally, a third contentious issue in the ethics of war regards situations of noncompliance. Do our moral obligations change when the enemy is not respecting the rules of war? In Walzer, this is the problem of reprisals, as reprisal violence or rights-violations are often the resort of the victim (see 1977: 207 ff). Are reprisals justified? How much risk should we require our soldiers to bear in these situations before loosening the constraints on their conduct? It is this very question that leads Walzer to formulate the doctrine of supreme emergency, discussed above. Walzer's answer, as we have seen, is that soldiers are not justified in violating the rules of war until the heavens really are about to fall (see, for example, 1977: 232, 251 ff).

However, as I argued above, one of the advantages of understanding supreme emergency through a consequentialist lens is that we would be able to make a broader range of important moral claims. One of those claims we might want to make—one plausible facet of *in bello* morality—is that soldiers can be justified in violating the rules of war when something *less* than a supreme emergency threatens them, especially when they are threatened with torture or enslavement on an individual level. Unlike Walzer, we cannot wait until the heavens are about to fall before we violate the rules of war.

Perhaps the solution has been developed within the rule-consequentialist literature instead. Noncompliance also poses a serious problem for rule-consequentialism, as the foundation of the theory is a stipulation about *what everyone else is doing*, despite the fact that it's rarely true that everyone else really is acting in the requisite way. Thus, it seems rule-consequentialism is especially vulnerable to the problem of free riders. Hooker's solution is remarkably permissive compared to Walzer's: we are no longer required to reciprocate toward those who are not obeying the rules of morality when it would require a significant personal sacrifice to comply (**?**: 125, 165 n. 10).

Hooker's requirement that the agent be faced with significant sacrifice is crucial, since it might vary in everyday morality whether we are required to reciprocate to free riders. In

times of war, however, the stakes are appreciably raised, and the enemy's noncompliance can immediately endanger the lives of soldiers. In situations where grave danger is commonplace, Hooker's criterion for violating the rules of war would be met almost routinely. If this were the case, solders would have license to disregard the rules of war at the first sight of noncompliance. This, clearly, is unacceptable.

There are other reasons to avoid Hooker's solution and, instead, strengthen the requirement that soldiers endure unfair treatment from the enemy before being licensed to retaliate with their own violations. For one, the rules of war might be thought of as more important than the rules of everyday morality because they have a greater potential to cause massive suffering. Even comparatively trivial rules of war regard the wellbeing and lives of many people. The choices before us are not whether to break a promise or to tell a lie; they are decisions about whether to bomb a civilian area, kill a prisoner of war, carry out mass arrests and detentions, or use morally problematic weapons such as white phosphorous. These decisions, which are far lesser violations than Walzer thinks a supreme emergency authorizes, are still clearly grave. We should hesitate to cast off the restraints of morality, as Hooker's solution would dictate, in the face of noncompliance.

Second, as mentioned above, this solution would amount to free license for soldiers and policymakers to do whatever they should want to the enemy at the first sight of noncompliance. And we should certainly want to avoid this situation. First, it threatens unrestrained massacre and reprisals. Second, it requires less of our soldiers than we think they ought to bear. We commonly expect our soldiers to respect the rules of war when fighting an enemy that *disregards* those rules, and even when it means they will bear a substantial burden as a result.

So, Walzer's approach is too stringent: we should endorse something less demanding. Hooker's solution is too liberal, and provides license for wanton violence in the face of a noncompliant enemy. We have to take a middle ground. We demand that our soldiers bear at least some hardship before loosening the restrictions on their combat. However, we

also realize they are justified in violating the rules of war in situations short of supreme emergency. We further recognize that greater dangers should license them to break more stringent restrictions.

As it is difficult to precisely quantify the value at stake in moral decisions, so it will be difficult here to draw an exact line for how much soldiers should have to endure before being allowed to break the rules of war. My intuition is that they ought to bear a substantial burden before that happens. Indeed, soldiers are thought to have to go to their own death before violating the most stringent rights of noncombatants.[16] We are justified in breaking the rules of war only when something terrible, outside the normal horrors of war, is almost certain to happen otherwise.

While definitive answers are still forthcoming, we can appreciate the clarity that this rule-consequentialist *in bello* code brings to the issue of noncompliance, as well as the other issues of warfare canvassed above. That this theory boasts a plausible consequentialist underpinning and does an excellent job of unifying and systematizing the rules of warfare, I believe, bodes well for its prospects for contributing to the military ethics tradition.

[16]For example, even if the only way a soldier may ensure her survival is by taking a human shield, she is still forbidden from doing so.

Bibliography

Adams, Robert Merrihew. 1976. "Motive utilitarianism." *The Journal of Philosophy* 73:467–481.

Anscombe, G. E. M. 1981. "Mr. Truman's Decree." In *The Collected Philosophical Papers of G. E. M. Anscombe*, volume 3 of *Ethics, Religion, and Politics*. Blackwell.

Baron, Marcia. 1987. "Kantian ethics and supererogation." *The Journal of Philosophy* 84:237–262.

Brandt, Richard. 1979. *A Theory of the Good and the Right*. Oxford University Press.

—. 1992. *Morality, utilitarianism, and rights*. Cambridge University Press.

Carson, Thomas. 1991. "A note on Hooker's rule consequentialism." *Mind* 100:117–121.

Dreier, Jamie. 2004. "Why Ethical Satisficing Makes Sense and Rational Satisficing Doesn't." In Michael Byron (ed.), *Satisficing and Maximizing*. Cambridge University Press.

Eggleston, Ben. 2007. "Conflicts of Rules in Hooker's rule-consequentialism." *Canadian Journal of Philosophy* .

—. 2013. "Rejecting the Publicity Condition: The Inevitability of Esoteric Morality." *The Philosophical Quarterly* 63:29–57.

Feinberg, Joel. 1961. "Supererogation and Rules." *Ethics* 71:276–288.

Gert, Bernard. 1988. *Morality*. Oxford University Press.

Guevara, Daniel. 1999. "The impossibility of supererogation in Kant's moral theory." *Philosophy and Phenomenological Research* 593–624.

Hare, R. M. 1965. *Freedom and Reason*. Oxford University Press.

—. 1979. "What is wrong with slavery." *Philosophy and Public Affairs* 8:103–121.

—. 1981. *Moral Thinking*. Oxford: Clarendon Press.

—. 1989a. "Principles." In *Essays in Ethical Theory*. Oxford: Clarendon Press.

—. 1989b. "The Structure of Ethics and Morals." In *Essays in Ethical Theory*. Oxford: Clarendon Press.

—. 1999a. "Objective Prescriptions." In *Objective Prescriptions and Other Essays*. Oxford University Press.

—. 1999b. "Prescriptivism." In *Objective Prescriptions and Other Essays*. Oxford University Press.

Harsanyi, John. 1977. "Rule Utilitarianism and Decision Theory." *Erkenntnis* 11:25–53.

Haydar, Bashshar. 2002. "Forced Supererogation and Deontological Restrictions." *The Journal of Value Inquiry* .

Hooker, Brad. 1990. "Rule-consequentialism." *Mind* 99:67–77.

—. 1991. "Rule-consequentialism and demandingness." *Mind* 100:269–276.

—. 1994. "Is rule-consequentialism a rubber duck?" *Analysis* 54:92–97.

—. 1996. "Ross-style pluralism versus rule-consequentialism." *Mind* 105:531–552.

—. 2000. *Ideal Code, Real World: A rule-consequentialist Theory of Morality*. Clarendon Press.

—. 2006. "Right, wrong, and rule-consequentialism." In Henry R. West (ed.), *The Blackwell Guide to MIll's Utilitarianism*, 233–248. Blackwell.

Hooker, Brad and Fletcher, Guy. 2008. "Variable versus Fixed-rate Rule-utilitarianism." *The Philosophical Quarterly* 58:344–352.

Horgan, Terence and Timmons, Mark. 2010. "Untying a Knot from the Inside Out: Reflections on the 'Paradox' of Supererogation." *Social Philosophy & Policy* 27:29–63.

Howard-Snyder, Frances. 1993. "Rule-consequentialism is a rubber duck." *American Philosophical Quarterly* 30:271–278.

Jamieson, Dale and Elliot, Robert. 2009. "Progressive Consequentialism." *Philosophical Perspectives* 23:241–251.

Jenkins, Ryan and Kahn, Leonard. In draft. "Epistemic variable-rate rule-utilitarianism."

Kagan, Shelly. 2000. "Evaluative focal points." In Dale Miller Brad Hooker, Elinor Mason (ed.), *Morality, rules, and consequences: a critical reader*, 134–155. Edinburgh University Press.

Kahn, Leonard. 2012. "Rule consequentialism and scope." *Ethical Theory and Moral Practice* 15:631–646.

—. 2013. "Rule-consequentialism and disasters." *Philosophical Studies* 162:219–236.

—. Unpublisheda. "Breaking the Code: A Problem for Rule-consequentialism."

—. Unpublishedb. "Rule consequentialism and compliance."

Lyons, David. 1965. *Forms and Limits of Utilitarianism*. Oxford University Press.

McNaughton, David and Rawling, Piers. 1991. "Agent-relativity and the doing-happening distinction." *Philosophical Studies* 63:167–185.

—. 1998. "On Defending Deontology." *Ratio* 11:37–54.

Mulgan, Tim. 2005. *The Demands of Consequentialism*. Oxford University Press.

Norcross, Alastair. 1997. "Consequentialism and commitment." *Pacific Philosophical Quarterly* 78:380–403.

—. 2005a. "Contextualism for consequentialists." *Acta Analytica* 20:80–90.

—. 2005b. "Reasons Without Demands: Rethinking Rightness." In James Dreier (ed.), *Blackwell Contemporary Debates in Moral Theory*. Wiley-Blackwell.

Oddie, Graham. 1998. "Moral Realism, Moral Relativism and Moral Rules (A Compatibility Argument)." *Synthese* 117:251–274.

Orend, Brian. 2006. *The Morality of War*. Broadview Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford University Press.

—. 2011. *On What Matters*. Oxford University Press.

Portmore, Douglas. 2007. "Consequentializing Moral Theories." *Pacific Philosophical Quarterly* 88:39–73.

Railton, Peter. 1984. "Alienation, consequentialism, and the demands of morality." *Philosophy and Public Affairs* 13:134–171.

Ridge, Michael. 2006. "Introducing Variable-Rate Rule-Utilitarianism." *The Philosophical Quarterly* 56:242–253.

Ross, W. D. 1930. *The Right and the Good*. Hackett Publishing Company.

Sartre, Jean-Paul. 1945. "Existentialism is a Humanism." Lecture.

Simon, Herbert. 1966. "A behavioral model of rational choice." *The Quarterly Journal of Economics* 69:99–118.

Singer, Peter. 1972. "Famine, affluence, and morality." *Philosophy and Public Affairs* 1:229–243.

—. 1982. *Practical Ethics*. Cambridge University Press.

—. 2010. *The Life You Can Save: How to do your part to end world poverty*. Random house Trade Paperbacks.

Slote, Michael. 1985. *Common-sense Morality and Consequentialism*. Routledge & Kegan Paul.

Smart, J. J. C. 1973. "Outline of a System of Utilitarian Ethics." In Bernard Williams J. J. C. Smart (ed.), *Utilitarianism: For and Against*. Cambridge University Press.

Suikkanen, Jussi. 2008. "A Dilemma for Rule-Consequentialism." *Philosophia* 36:141–150.

Taylor, Charles. 1997. *Philosophical Arguments*. Harvard University Press.

Urmson, J. O. 1958. "Saints and Heroes." In *Essays in Moral Philosophy*. University of Washington Press.

Walzer, Michael. 1977. *Just and Unjust Wars*. Basic Books, fourth edition.

—. 2004. *Arguing About War*. Yale University Press.

Woodard, Christopher. 2008. "A New Argument Against Rule Consequentialism." *Ethical Theory and Moral Practice* 11:247–261.

—. 2013. "The Common Structure of Kantianism and Act-Utilitarianism." *Utilitas* 25:246–265.

Appendix A: Individual rule-consequentialism and collapse

Rule-consequentialism (RC) determines the rightness of actions by reference to an ideal moral code (IMC) of rules which applies to all (or almost all) agents. This is the case in RC's canonical formulations, for example, Brandt (1979), Harsanyi (1977), Hooker (2000), and Parfit (2011). Relativized forms of rule consequentialism, which specify different ideal moral codes for different social groups, have been considered before[17], yet they have not received serious attention until recently.[18] Leonard Kahn's recent article, "Rule-consequentialism and scope" (2012) is one such discussion that provides a powerful argument for relativizing the ideal moral code.

Proponents of rule-consequentialism often appeal to either the maximization of aggregate welfare or coherence with our considered moral intuitions (or both) to support their view. Kahn's argument appeals to rule-consequentialists who endorse rule-consequentialism for either of these reasons. First, he argues that relativizing the ideal code to social groups would plausibly make some people better off while making no one worse off. Second, he notes that this relativized view is thus coherent with our intuitions regarding the desirability of costless benefits, i.e. it better passes the test of reflective equilibrium.

Kahn lays out three possibilities for a version of relativized rule-consequentialism:

[17]See for example (Brandt, 1979: 194) and Hooker (Hooker, 2000: 189).

[18]See especially Kahn (2012), the focus of this note. See Oddie (1998) for an earlier discussion that is tangential to the points made here.

> *Weak RC*. For each relevant group there is an IMC, where G is a group with regard to some set S of agents if and only if G is composed of two or more agents who form a proper subset of S. (a.k.a, the Weak Thesis) (2012: 632)
>
> *Moderate Relative RC*. Some morally ideal codes hold for groups of agents while others hold for a single agent. (a.k.a., the Moderate Thesis) (2012: 633)
>
> *Strong Relative RC*. For each agent there is exactly one IMC (a.k.a., the Strong Thesis) (2012: 633)

The most radical of these is the strong thesis: the view that every individual ought to internalize an idiosyncratic moral code. Kahn ultimately endorses the moderate thesis, as he believes that rule-consequentialists have good reason to "stop short," as it were, of endorsing a unique moral code for each person. Still, Kahn spends some time considering what this radical view would entail. In his discussion that Kahn leaves out important comparisons between individual RC and other popular moral views. Here, I explore these overlooked comparisons.

First, a more complete explication of Kahn's Strong Relative RC, which I will call Individual RC:

> *Strong Relative RC* (a.k.a., *Individual RC* or simply IRC). It is wrong for agent A to perform an action if is forbidden by the code of rules whose internalization by A would lead to maximum expected consequences in terms of wellbeing, impartially considered, other things being equal.

Kahn considers the possibility that Individual RC collapses into equivalence with other moral theories. Kahn entertains these suggestions for the sake of clarity, he says, but there is more at stake in this discussion: If IRC collapsed into equivalence with some other moral theory, it would presumably inherit the vices (and virtues) of that other theory. That is, IRC would be problematic if it collapsed into equivalence with some other objectionable view. If it could be shown to do just that, then rule-consequentialists would have an additional reason,

appealing to reflective equilibrium, to reject IRC. On the other hand, if IRC collapsed into some eminently plausible view, then it would benefit by association. What we are examining here is whether IRC stands or falls with any other better-known theory in ethics. The answer is ultimately No, and we must judge this nascent moral theory on its unique merits.

Kahn considers the possibility that IRC collapses into equivalence with maximizing act-consequentialism[19] and ethical egoism, and shows that it does not collapse into equivalence with either (2012: 642). Here, I examine whether IRC collapses into equivalence with two other views, sophisticated consequentialism and motive utilitarianism, to which IRC bears a striking resemblance. I will show that it collapses into equivalence with neither.

First, note we might mean at least three things by *equivalence* in this context.[20] This distinction allows for greater nuance in our discussion than others have given. Lyons (1965) provides the seminal discussion of equivalence within the consequentialist tradition. There, Lyons says two principles are "extensionally equivalent" if they "always yield substantively identical (equivalent) judgments with respect to particular acts" (Lyons, 1965: 29). Notice, though, that two theories may yield equivalent judgments about which acts are right and wrong, but disagree about an agent's praiseworthiness, or about what ought to be done in order to successfully follow a theory. Hooker is not concerned, either, with making such fine-grained distinctions when he argues that rule-consequentialism does not collapse into equivalence with maximizing act-consequentialism (2000: 93–99). Yet it is possible, as I show below, that two theories may be equivalent in one of these respects and inequivalent in another.

[19]See Lyons (1965). Note that this is separate from the charge that rule-consequentialism in its more familiar universal formulation collapses into extensional equivalence with maximizing act-consequentialism. That is the standard "collapse objection," which can be found in Smart (1973) among others.

[20]Note that Kahn does not mention *equivalence* in his original article, but rather *collapse*, using the shorthand common in discussions of the collapse objection to rule-consequentialism.

Behavioral equivalence. Two theories are behaviorally equivalent if and only if an agent who successfully follows the one performs all the same actions[21] as an agent who successfully follows the other.

Rightness equivalence.[22] Two theories are rightness equivalent if and only if they agree on the rightness and wrongness of all actions.

Praise equivalence. Two theories are praise equivalent if and only if they agree on the praise- and blameworthiness of agents with regard to all actions.

Finally, consider the following complex judgment of equivalence:

Content equivalence. Two theories are content equivalent if and only if they are behaviorally equivalent, rightness equivalent, and praise equivalent.

According to IRC, each person ought to internalize the code of rules that, given her peculiar psychological makeup, will maximize expected utility, taking into account the code's internalization costs. Following Hooker (2000), let's suppose that to internalize a rule involves forming the associated motivations to comply with the rule, but also involves

> having sensitivities, emotions, and beliefs—indeed a particular cast of character and conscience. If you accept a rule against stealing, you will be motivated not to steal simply because it is stealing... You will also be disposed to feel guilty if you steal, disposed to resent stealing by other people, and disposed to blame them for it. (2000: 91)

On its face, IRC bears a resemblance to some other forms of consequentialism: sophisticated consequentialism and motive utilitarianism. IRC collapses into behavioral equivalence

[21]For simplicity I will not distinguish between acts and actions. I do take an agent's mental states such as intentions to be partly constitutive of an action.

[22]Woodard (2013) calls this "deontic equivalence." In reference to Portmore's work on consequentializing (2007), Woodard says: "To consequentialize a theory, one must devise a theory of value that, when combined with act-consequentialism, yields the target theory's deontic verdicts in all cases" (2013: 261). Thus, what Woodard has in mind by "deontic equivalence" (2013: 262) is a necessary—but not the sufficient—condition for what I here call *content equivalence*.

with the first of these theories and a modified version of the second, but does not collapse into content equivalence with either.

According to sophisticated consequentialism, an agent ought to cultivate the set of dispositions that will allow her to maximize her expected contribution to wellbeing, impartially considered, over the course of her life.[23] So far, this is just what IRC requires, so sophisticated consequentialism and IRC are behaviorally equivalent. However, the two theories differ significantly in their criteria of rightness and in their assignments of praise and blame, and so are not content equivalent. Because sophisticated consequentialism is a species of maximizing act-consequentialism, the right action in any particular instance is the action that contributes the maximum expected amount of wellbeing to the world, even though this is frequently not the action supported by the ideal set of dispositions. In this somewhat bizarre fashion, even the scrupulous sophisticated consequentialist must accept that she frequently, or perhaps always, performs the wrong action. She merely denies that this is blameworthy. According to IRC, though, the right action is the action that follows from the ideal set of dispositions, and acting in accordance with those dispositions is praiseworthy. Thus, IRC does not collapse into content equivalence with sophisticated consequentialism.

Motive utilitarianism receives its seminal explication in Adams (1976). On motive utilitarianism, the morally perfect person "would have the most useful among the patterns of motivation that are causally possible for human beings" (Adams, 1976: 470), where a person's pattern of motivation is the sum total of "wants and desires, considered as giving rise, or tending to give rise to actions" (Adams, 1976: 467). If *having internalized a rule* for IRC means the same as *having a motive* for motive utilitarianism, then IRC seems to collapse into behavioral equivalence with motive utilitarianism.

At first it seems that internalizing the individual ideal moral code of rules might result in the same conscience as would becoming the ideal motive utilitarian agent, since both

[23]See Railton (1984) for sophisticated consequentialism's canonical formulation and defense. See Norcross (1997) for further elaboration.

theories recommend the set of dispositions that would be expected to maximize wellbeing, impartially considered, over the course of one's life. In fact, it would not. Because these theories are not behaviorally equivalent, they cannot be content equivalent, either. They are not behaviorally equivalent because while both theories seem to recommend the same set of motivations, they do not. In evaluating competing codes of rules, i.e. competing sets of dispositions, rule consequentialists take into account the internalization costs of each code (Hooker, 2000: 78–80). Thus, while it might maximize expected consequences to *have internalized* some code of rules—perhaps the set of motives that Adams thinks optimific—that code might have a prohibitively high cost *to internalize* and so would be ruled out by rule-consequentialism. Adams makes no such allowance for the internalization costs of sets of dispositions. Thus, some sets of dispositions are open to him that are ruled out by rule-consequentialism.

However, there is one modification Adams could make to motive utilitarianism such that the two theories would be behaviorally equivalent. Let's call "wants and desires, considered as giving rise, or tending to give rise to actions" *first-order* motives, and let's call motives regarding the formation of first-order motives *second-order* motives. Thus, I may have a first-order motive to eat meat, and a second-order motive to reform that particular motivation in order to become a vegetarian. Suppose the ideal motive utilitarian agent has a second-order motive which moves him to bring his present set of motives into line with the ideal set recommended by motive utilitarianism. Surely the inclusion of this second-order motive in the motive utilitarian calculus will change the set of all motives, first- and second-order, that would best maximize expected utility, since acting on his second-order motive will involve shouldering certain burdens, exerting efforts, enduring psychological strain, etc., in the pursuit of perfecting his first-order motives. Hence, acting on this second-order motive carries a cost analogous to the internalization cost of the ideal code for rule-consequentialism. In this case, IRC and motive utilitarianism move closer together. In fact, I suspect the set of first-order motives each theory would recommend would be identical, given this modification,

and so would be the actions each would recommend. Thus the theories would be behaviorally equivalent.

It is worth noting additionally that, for Hooker, an agent need not act from a rule having been internalized, but need only act in accordance with the ideal set of rules, in order to perform a right action. Not so on Adams' theory. Motive utilitarianism is a theory about which *motives* an agent ought to have and not a theory of which *actions* are right (Adams, 1976: 474). If Adams requires certain mental states for an action to be right, for example, then the theories will not be content equivalent. But he is silent on the matter and so we cannot know whether IRC and our modified motive utilitarianism are content-equivalent.

Individual rule-consequentialism does not collapse into equivalence with maximizing act-consequentialism or egoism, as was previously shown. Here, it was shown that individual rule-consequentialism does not collapse into sophisticated consequentialism or motive utilitarianism, either, two theories with which it bears a remarkable superficial similarity. Individual rule-consequentialism must be evaluated on its own merits, then. Rule-consequentialists of some stripes, for example, those who argue from a kind of universalization or contractualism, will not be moved by the arguments for IRC. But this new theory seems to satisfy our desires both to make the world a better place and to cohere with our considered moral intuitions, in particular, our intuitions regarding costless benefits. These are two notable advantages of the theory. An all things considered judgment of the plausibility of individual rule-consequentialism would require a much broader investigation, which must be left for another time.

Appendix B:

An argument against communitarianism

In the hands of Walzer, supreme emergency is justified only by a grave threat to a political community *as such*. In turn, this fits well within Walzer's general communitarian framework. However, in the fourth chapter of this dissertation, I developed a rule-consequentialist account of just war theory. If we are to justify the resort to supreme emergency within the consequentialist tradition rather than the communitarian one, we must remove the emphasis that Walzer places on the political community itself. To bolster those arguments, I give several reasons in this appendix to dismiss the communitarian claim that a political community has value in itself. I take Charles Taylor's argument (1997) to be characteristic of those made by philosophical communitarians.

Taylor's argument, briefly, is this:

(1) A particular culture is essential to the enjoyment of some irreducible goods, such as *fulfillment* or *freedom*. That culture is the necessary background condition against which obtaining these goods becomes possible.

Taylor argues that, for certain goods, it is *incoherent* to say a person is enjoying those goods without referencing her cultural context. "Thoughts presuppose and require a background of meanings to be the particular kind of thoughts they are. But the terms 'presup-

pose' and 'require' in the previous sentence point to a peculiarly strong relation... Certain thoughts are impossible in certain circumstances" (1997: 132). Taylor adds, "As individuals we value certain things; we find certain fulfillments good, certain experiences satisfying, certain outcomes positive. But these things can only be good in that certain way, or satisfying or positive after their particular fashion, because of the background understanding developed in our culture" (1997: 136).

(2) If x is necessary for the enjoyment of some intrinsic good y, then x is itself intrinsically good.

Taylor makes this rather surprising claim at (1997: 137). He says that if x is "essentially linked" to y, or "analytically undecomposable" from y, then x is not only instrumentally good, but is itself *intrinsically* good. He continues, "[A culture] can't be distinguished from them [the individual goods it makes possible] as their merely contingent condition, something they could in principle exist without... Consequently, it is hard to see how we could deny it [the culture] the title of good, not just in some weakened, instrumental sense... but as intrinsically good" (1997: 137).

$\therefore$ (3) A certain culture, over and above the individuals that live in it, is itself an intrinsic good. (MP 2 & 3)

There are a number of criticisms we could make of this argument. First, (2) above seems false. Being able to show that my explanation for why y is intrinsically good must refer to some x, i.e. y is analytically undecomposable from x, does not seem to be sufficient to show that x is also intrinsically good. Suppose I particularly enjoy going for lazy Sunday drives around my neighborhood. Suppose further that I receive a particular kind of pleasure from this, a kind of pleasure that could not be secured any other way. My Sunday drives are certainly a good for me. But according to Taylor, my neighborhood itself is also a good. It is certainly plausible to say here that my neighborhood has some instrumental value for me,

but Taylor goes further: he has to say that my neighborhood is an intrinsic good because it is essential to the securement of some further good—after all, I could not get *this* particular enjoyment by driving around some *other* neighborhood. But that seems to mistake the idea of an intrinsic good. It is coherent to say instead that, even if some good x is the only way of securing good y, that x is merely an instrumental good.

Regarding (1) above, it seems plausible to say that some culture provides the linguistic content of a concept like *freedom*, for example. However, it seems false that having any such linguistic content is necessary to enjoying freedom, i.e. that it is necessary to the enjoyment of being free. More plausible is the epistemic claim that, while I may enjoy x, I get *even more* pleasure by *recognizing* that I am enjoying x. But that recognition is itself not an essential part of the original enjoyment. Before *freedom* took on the sense that it did with the rise of liberal democratic states in the 18th and 19th centuries, it seems that people could have still taken pleasure in the possession of some autonomy, i.e., in the ability to choose to live in ways consistent with their own values and desires. This is true even if the pleasure of such an experience were inexplicable or ineffable. It seems a similar argument could be given regarding the most important kinds of "irreducible" goods in our lives: personal autonomy, political self-determination, close, loving relationships, pleasure, fulfillment, etc.

Finally, with regard to just war theory in particular, notice also that Taylor's argument is only successful if consequentialism is committed to what he calls *atomism*. His argument above is meant to show that atomism leaves something out, since it cannot account for the additional value located in the ongoingness of a community. There are two ways to understand Taylor's statement: first, as asserting that consequentialist theories only count the pleasure and pain that sentient creatures enjoy. If this were true, consequentialism could not account for an extra good that is not located in any particular individual, i.e. the good constituted by the *ongoingness* of a political community. This is usually, but not always, true of consequentialism. In fact, it could locate value anywhere, not just in sentient creatures.

A version of consequentialism inspired by G. E. Moore might locate value in the beauty of a rose thousands of light-years away from here.

Second, we might understand Taylor's atomism as a rejection of organic unities. Thus, atomism might be the view that that moral value aggregates "simply," such that nothing can ever be more valuable than the sum of its parts. Again we can turn to G. E. Moore to show that Taylor is wrong: one can be a consequentialist and embrace organic unities. There is nothing in the defeat of atomism, if Taylor really has defeated atomism, that entails the defeat of consequentialism. Consequentialists usually locate value in the wellbeing of conscious creatures, but they could also ascribe value to the ongoingness of a culture. I have been at pains to undermine the claim that there exists some extra value located in political communities as such, since this claim is essential to one version of the doctrine of supreme emergency within just war theory. If I have failed, and there is some irreducible value to be found in political communities as such, consequentialism can still make sense of such value, clearing the way for a consequentialist account of supreme emergency.