Second order consequentialism: a defense

SECOND ORDER CONSEQUENTIALISM: A DEFENSE

By

Nihar Nandan Nilekani

A Thesis submitted in partial fulfillment

of the requirements of the degree of

Doctor of Philosophy

Department of Philosophy

Birkbeck, University of London

I, Nihar Nandan Nilekani, declare that this thesis has been composed solely by myself. All work presented is my own except where explicitly stated otherwise and referenced. It has not been submitted for any previous application for a degree.

ABSTRACT

This thesis is a defense of **Second Order Consequentialism (SOC)**. Whereas first order consequentialism is the claim that we should perform the action that results in the best consequences, SOC is the claim that we should adopt the moral theory for which it is true that adopting it would have the best consequences. I argue that this form of consequentialism has many of the traditional virtues of consequentialism, but by virtue of its indirectness it conflicts less with our intuitions and produces more desirable outcomes at the level of moral decision making.

I argue that we have good reasons to make a distinction between first and second order theories, independent of any problems with consequentialism. I further argue that certain ethical problems – most notably the problem of demandingness and the concept of threshold deontology – cannot be properly addressed without a second order theory of some kind. I will then apply SOC to these areas to demonstrate how a second order theory can be used to fruitfully address problems like these. Finally, I will explore the idea of first order pluralism with a unifying second order theory.

Second-order theories in general provide us with a framework by which we can interrogate our intuitions, particularly those that we have reason to think are in some way suspect or biased. Further, second order theories can be more easily made sensitive to changed circumstances or relevant information, without requiring awkward additions or clauses attached to the theory, as is often the case with first-order theories. Many of these are virtues of all second-order theories, including nonconsequentialist ones, but second order consequentialism is particularly promising due to its relative simplicity, and I think worth considering first before we jump to more, perhaps unnecessarily complicated, second order theories.

Contents

PART 1: LAYING THE FOUNDATIONS 5
Chapter 1: Introduction 5
Chapter 2: The Role of our Intuitions 13
Chapter 3: Why Consequentialism? 34
Chapter 4: Why Second Order? 47
Chapter 5: What is Second Order Consequentialism? 67
PART 2: DEMANDINGNESS 88
Chapter 6: The Demandingness Objection 89
Chapter 7: Traditional Answers to the Demandingness Objection 109
Chapter 8: Second Order Consequentialism and Demandingness 130
PART 3: BLAMEWORTHINESS 148
Chapter 9: Defending My Theory of Blameworthiness 149
Chapter 10: Second Order Consequentialism and Collective Responsibility 181
PART 4: FIRST ORDER PLURALISM 206
Chapter 11: Catastrophes, Thresholds and Vagueness 207
Chapter 12: Circumstantial Moral Principles 227
Chapter 13: Relationships, Honoring, and Virtues 245
PART 5: CONCLUSION 264
REFERENCES 268

PART 1: LAYING THE FOUNDATIONS

Chapter 1: Introduction

The thesis

This thesis is a defense of **second order consequentialism**: the claim that we should adopt and internalize that set of moral principles for which it is true that adopting them would lead to the best consequences in the long run. This is in contrast to **first order consequentialism**, which would have us do whatever action leads to the best consequences. A first order moral theory is one that tells us what to do when we are faced with a decision. But a second order theory doesn't apply to actions directly; rather, it is a framework for evaluating and justifying first order theories. It tells us *why* we should use the decision making criteria of our first order theory. In this thesis I want to claim that second order consequentialism (sometimes abbreviated **SOC**) is both a very promising example of a second order theory and a way to gain most of the benefits of consequentialism while avoiding many of its problems.

Taking an indirect approach is something consequentialists have done for a long time[1], such as the two-level consequentialism of R.M. Hare (1981). However, my theory makes a much sharper division than any of the preceding approaches between what Kagan (1998) calls the factoral and foundational levels of moral theory, which I call the first and second orders. In this sense it is most similar to a rule consequentialist theory such as that of Hooker (2002), with the main difference being that I am not committed to the idea that the correct first order theory is a set of rules, or even to there being only one correct first order theory.

[1] (Bentham J. , 1840) (Mill, 1861)

I believe that this form of consequentialism has many of the traditional virtues of consequentialism, but by virtue of its indirectness it conflicts less with our intuitions and produces more desirable outcomes at the level of moral decision making. Furthermore, a second order moral theory provides us with a framework by which we can interrogate our intuitions, particularly those that we have reason to think are in some way suspect or biased. Finally, second order theories can be more easily made sensitive to changed circumstances or relevant information, without requiring awkward additions or clauses attached to the theory, as is often the case with first-order theories. Many of these virtues, I will argue, are common to all second-order theories, including nonconsequentialist ones. However I will also argue that second-order consequentialism is a particularly promising approach, and due to its relative simplicity is worth considering first before we jump to more, perhaps unnecessarily complicated, second order theories.

My overall goal in this dissertation is thus twofold. First, I will argue that some form of second order moral theory is necessary for dealing with many problems of ethics, including the problem of demandingness and puzzles that arise when dealing with thresholds and moral dilemmas. This is particularly true when we have good reason to think our intuitions might be unreliable, as we do in these cases. A second order theory gives us the means to criticize and interrogate our intuitions. Secondly, I will argue that SOC is a very promising example of such a moral theory. Not only does it give us ways of dealing with these problems, I will also argue that it is a generally fruitful approach to moral theorizing. SOC has many of the main benefits of consequentialism, such as relative simplicity and unity of explanation. By contrast, many objections to classical utilitarianism and other first order consequentialist theories do not apply to SOC.

The structure of the argument: theory holism and plausibility points

My aim in this dissertation is to promote what I believe is a very fruitful ethical theory, and I believe that is best served by showing that the approach is useful in a wide variety of cases rather than focusing on a single one. As such, this dissertation takes a wide overview showing how my theory applies in diverse cases rather than focusing on a single application where I explore every implication and defend against every imaginable criticism. Such an exercise can be fruitful, but focusing on a singular argument runs the risk of causing one to lose track of the larger context. In addition, I am skeptical of singular knockdown arguments that can by themselves cause one to adopt or abandon a philosophical theory. I do not deem theories implausible because of some single counterargument or troubling example. Even with theories I disagree with, I can usually come up with defenses for them against any *single* argument. If I don't think a theory or approach is viable it is either because it has a large number of separate problems or because I find the initial reasons given in its favor uncompelling, rather than because it has some single critical flaw. And as for theories I am convinced to believe in because of some single argument, I can think of none. We adopt, rather, those theories that we have the most reason to believe in as a *whole*.

In his defense of Robust Moral Realism, David Enoch (2011) introduces the idea of 'plausibility points' to describe how he goes about giving this kind of holistic argument for his theory. Enoch concedes that moral realism, especially of the non-naturalistic kind, has many serious problems as a metaethical theory. However, he also asserts that all rival metaethical theories have similarly serious problems – there is no truly unproblematic metaethics. If we have reasons for believing in a theory, that gives that theory plausibility points; arguments against that theory cause it to lose plausibility points. When defending his theory, Enoch is not trying to

defend his theory from all possible criticism, which would be as futile as producing an unbeatable criticism. Rather, he is trying to *minimize* the loss of plausibility points, while acknowledging that some amount will be lost if the criticism has any force to it. When giving arguments in favor of his theory, he is trying to maximize the gain in plausibility points he can get out of any particular argument. In the end, the theory that has the greatest net amount of plausibility points as a whole, out of all rival theories, is the one we should believe in.

This is broadly the approach I am taking as well. My theory unquestionably has its flaws and weak points. While it is an attempt to produce a version of consequentialism that is more in line with our intuitions, it also asks us to revise those intuitions quite heavily, as I will detail in later chapters. But so too will any theory that attempts to criticize and analyze our intuitive judgements, which is of course most of them. The only type of theory that is not in some sense revisionary is one that does not attempt to critique our intuitions at all, and such a theory is implausible in its own right. My theory sometimes appears strange and occasionally makes some rather startling claims: for example, that our judgments about whether or not an action is blameworthy are partially justified on the basis of the consequences of such judgements rather than just the consequences of that action itself. But it is my belief that we have good reasons to adopt SOC when the theory is viewed as a whole, despite individually implausible elements. The way this dissertation is organized reflects this holistic approach to philosophical theories.

The structure of the dissertation

This dissertation is divided into several main parts, each containing multiple chapters. This first part is mainly introductory and attempts to lay the groundwork for my theory, while the rest of the dissertation is focused on applications of the theory to several areas of ethical theory where I think it is useful and productive: namely, the problem of demandingness, a variant

approach to blameworthiness and responsibility, and the possibility that we might adopt a pluralism about first order moral theories.

In this chapter I explain why and how I am arguing for the theory in the way I am. One of my main motivations for developing SOC is to produce a version of consequentialism that has less nonintuitive conclusions than most first order consequentialisms do, but I also want a theory that can tell us how to revise our intuitions if necessary. Thus, in the second chapter I will explain what role our intuitions have to play in moral discourse. The next two chapters are about motivating SOC. Chapter 3 is where I explain what makes consequentialism appealing as a moral theory and why I think it is an approach to ethics worth investigating despite its seeming flaws. Chapter 4 is where I go into detail about the difference between the first and second orders of moral theories and why I think we need second order moral theories. These two lines of argument to a certain extent stand on their individual merits: that is, I think we have very good reasons to embrace consequentialism and also think, separately, that we have good reasons to develop some kind of explicitly second order moral theory. But of course, combining those two claims gives us reason to adopt second order consequentialism. Chapter 5, therefore, is an in-depth explanation of SOC, including what I mean by 'adopting and internalizing' a first order theory, as well as what makes SOC different from rule consequentialism, with which it otherwise shares a lot of similarities.

The next part is about the problem of Demandingness, generally regarded as one of the biggest problems that consequentialist ethical theories have to deal with[2]. In chapter 6, I will lay out the problem and why anyone developing any kind of consequentialist theory needs to grapple

[2] see, e.g. Mulgan (2001)

with it. But I will also argue that it is not a problem for consequentialists specifically, as it is sometimes portrayed, but rather that the general question of how demanding morality should be is one that every moral theory must provide an answer to. In Chapter 7 I will go over some consequentialist answers to the problem of Demandingness and why I think they don't work: arguing that they and indeed any answers to the problem are incomplete unless they provide a justification for what the right level of demandingness is. In chapter 8 I will give my own answer to the objection, explaining how taking a second order consequentialist approach to the problem gives us a framework to determine in a non-arbitrary way how demanding our first order theory ought to be; in the process, I will develop a new way of thinking about blameworthiness.

The next part is about defending and exploring that notion of blameworthiness: that we ought to adopt the standards of blameworthiness which, if adopted, would lead to the best consequences in the long run. Chapter 9 is a long defense of this theory of blameworthiness against many possible objections, such as those arguing that it allows for cases where we blame people in ways that are clearly wrong if doing so would lead to good consequences. I will argue that this is not true once we understand that my theory asks that we adopt blaming practices rather than justifying particular instances of blaming – and only if they lead to the best consequences *overall* and not merely good consequences in some circumstances. In Chapter 10, I will apply my theory of blameworthiness to cases of group or collective responsibility, arguing that doing so allows us to avoid an otherwise extremely troubling dilemma: that any theory that charges us with doing something about collective problems like pollution is too demanding on the individual level, and conversely that any theory that is not too demanding on individuals is too permissive when it comes to these problems. I will argue that we can avoid this dilemma by embracing a notion of group action and using SOC to create a priority of such actions on the

basis of effectiveness, such that we are blameworthy for failing to be part of the right kind of group action but not for simply being less than perfect in our individual contributions.

The final part explores the idea of circumstantial moral theories: that a second order moral theory might justify not one first order theory but instead multiple different first order theories that apply to different circumstances. In Chapter 11 I examine the similar idea of threshold deontology, which is basically that deontological constraints (such as the prohibition against doing harm) cease to apply in exceptional or catastrophic circumstances where a very large number of lives is at stake. I will argue that threshold deontology has a deep arbitrariness problem due to lacking any sort of second order theory that might justify and locate the threshold. In Chapter 12 I will consider the idea of thresholds through the lens of SOC, and argue that we both do and should apply different sorts of moral standards to small scale and large scale cases. In Chapter 13 I will apply this pluralism about first order theories to a different kind of problem: that applying our ordinary moral theories to our close personal relationships seems to involve us interacting with those close to us for entirely the wrong reasons. I will argue that special obligations should not be regarded as some sort of exception carved out of a more general moral theory, but that we need a different kind of moral theory to explain how we should act with those close to us: that different first order theories govern our interactions with strangers and with friends.

Finally, in the conclusion I will make my final case for second order consequentialism, once its implications and benefits (as well as drawbacks) have been explored. As I said, I believe that theories must be evaluated holistically and not on the basis of an individual argument, and this is especially true since I am advocating for a broad approach to ethical theories more than a complete single theory – I am silent about my theory of the good, for example. It is my hope that

by the end of the dissertation, it will be obvious that we need some form of second order moral theory to be able to truly address certain ethical problems such as demandingness and thresholds, and that second order consequentialism is a promising second order theory that provides us with many of the benefits of consequentialism without its downsides.

Chapter 2: The Role of our Intuitions

Introduction

In this chapter I describe the attitude I think we ought to take about our intuitions when we do moral theorizing. A major goal of mine is to try and produce a version of simple maximizing consequentialism that is overall in harmony with all our most important intuitions. In fact, I wish to go a step further and produce a moral theory that is capable of analyzing our intuitions and interrogating them, allowing us to reexamine and reevaluate them. Overall I believe that our moral intuitions are like experimental results in science: they are a necessary starting point and test for our moral theories, but they are not themselves immune to recontextualization and criticism. But I do not believe that we should discard them, or rather I believe we should only discard particular intuitions after careful consideration and intuitions generally never. This attitude puts me at odds to many consequentialists, who are often generally skeptical of our intuitions, so in this section I will defend our intuitions against the most common criticisms of their reliability.

Setting up the problem

My attempt to analyze the role of intuitions in moral theorizing is particularly important because my overall project is a consequentialist one. There is a rich history of objections to consequentialism on the basis of it conflicting strongly with our deeply held intuitions. Indeed, it can be fairly said that the main overall objection to consequentialism is that it too often has nonintuitive conclusions. On the other side, many consequentialists, such as Peter Singer (2005) and Peter Unger (1996), argue for various reasons that our intuitions are not a generally reliable guide to moral theorizing, and thus that their conflict with consequentialist reasoning is not a

mark against the latter. They engage in deflationary arguments about our intuitions, arguing for one reason or another that our intuitions are unreliable or that we have reason to believe they do not track moral truth, and so are skeptical of them. To a large extent my overall project of second-order consequentialism is an attempt to try and reconcile our intuitions within a purely consequentialist framework, so I am trying to sidestep this conflict more than I am trying to argue for one side or the other. But for that very reason, it is essential that I lay out what I believe to be the role of our intuitions in moral theorizing.

Deflationary arguments about our intuitions to defend consequentialism often come up during discussions of demandingness, so I will here borrow Tim Mulgan's framing of it from his book *The Demands of Consequentialism*, as I feel it is a useful starting point. (2001). Although his discussion of this argument is brief[3], it is also insightful. Mulgan first describes the general strategy of what he calls the Extremist. The Extremist starts with a general moral principle that they claim to be nearly self-evident - e.g. Singer (1973) argues that we have an obligation to render aid to others if we only need to sacrifice something of lesser value - that motivates the consequentialist position. The Extremist justifies this principle with various arguments, usually including thought experiments. The next stage of the Extremist argument is to systematically argue against any limitation on the *scope* of this principle: arguing for instance that distance is morally irrelevant, as is the level of sacrifice. Since the original principle is highly demanding, the Extremist argues that so is morality. In thusly arguing against limitations, the Extremist must argue against very strong intuitions that we have. The Extremist generally does so by providing a deflationary account of the origin of our moral intuitions and arguing that any intuition that is not

[3] I will discuss Mulgan's other arguments in more detail later throughout Chapter 5, during my own discussion of demandingness

insufficiently justified must be discarded. For instance, Unger calls his account 'Liberationist', by which he means that it is meant to liberate us of the (in his view incorrect) notion that our intuitions in particular cases are reflective of our deepest moral commitments (Unger, 1996, p. Section 1.3). Another example of this deflationary account of our intuitions is Singer's paper on Ethics and Intuitions (2005). In it, Singer gives a brief account of the current state of moral psychology at the time, and argues that its normative implications are that we should abandon our reliance on intuitions and instead focus on our moral reasoning.

When these consequentialists cast doubt on our intuitions, they do not appear to merely mean our unreflective thoughts on situations, i.e. what we might call our 'gut feelings'. Casting doubt on such 'pre-theoretic' intuitions is relatively easy, but also does not get Denialists very much. Consequentialism also often clashes with our intuitions even after they have gone through a process of reflection – after they have survived discussion, attempts to put ourselves in the shoes of others or take an impartial view, etc. Consequentialists cast doubt on these all-things-considered intuitions as well. They argue by various means – which we will discuss in detail below – that even our reflective intuitions are the result of processes that are biased or flawed in certain ways, which means we should be greatly skeptical of the results. But this argument opens them up to a very hard to refute counter-argument, as shown by Mulgan's reply.

Mulgan's response to the Extremist's deflationary argument reminds me strongly of an argument in a different realm of philosophy entirely: Thomas Reid's argument against the skeptic of the external world. To the empiricist who doubts that there is a physical world beyond our senses, Reid points out that the skeptic still must have principles which they hold to be self-evident, such as the law of non-contradiction, for which they have no further justification. But our strong perception that there exists a physical world is no less self-evident: "Why sir," Says

Reid "should I believe the faculty of reason more than that of perception? —they came both out of the same shop, and were made by the same artist; and if he puts one piece of false ware into my hands, what should hinder him from putting another?" (Reid & Brooks (ed.), 1764/1997, pp. 6.20, 168-169) Were we to deny the validity of our intuitions in one case, we should do so in all cases - but to do so is to leave us with a skepticism so total that it becomes meaningless.

Similarly, Mulgan says to the Extremist: can your own moral principle stand up to the barrage of doubt you throw on all our *other* intuitions? The Extremist cannot throw doubt on all our moral intuitions, as without any moral intuitions whatsoever one cannot have a moral theory at all. Yet Extremists generally do not provide an argument for their consequentialist starting point that can hold up to the arguments they later use to demolish any limitations on that starting point, as their line of argument is primarily negative rather than positive in nature. Singer attempts to contrast reliance on our moral intuitions with an ethics instead grounded in our sense of reason, and attempts to avoid the problem that way. But this is *precisely* the kind of argument that Reidian arguments are meant to counter: our moral reasoning ability is just as much a product of our evolutionary history as our moral intuitions. Both our ability to generate reflective intuitions and our abstract reasoning abilities came out of the same shop and were made by the same artist. If we are to cast doubt on one, we must cast doubt on the other on the same grounds.

In response to this line of attack, Extremists/Liberationists must firstly identify some way of distinguishing our sources of moral knowledge from each other by defining more precisely 'intuitions' and 'reasoning'. But it is not clear that this can be done in a systematic way: in particular, it is not clear that our moral reasoning is meaningfully *different* from the above process of reflecting on our intuitions. Denialists argue that "utilitarians are at an advantage over those who hold moral views that are based on our commonly accepted moral rules or intuitions."

(Singer & de Lazari-Radek, 2017, p. 286) because they only rely on broad principles like "maximize the utility of all sentient beings" which can be arrived at purely via our rationality (*Ibid.*). But that is far from the only moral principle that can be arrived at via such means! Kant believed his theory was the result of pure reason as well. Many of our 'commonly accepted moral rules or intuitions' – those that undergo reflection and testing at least – are also the result of the same sorts of processes. By drawing a distinction between intuitions and reasoning Denialists are trying to make sharp a line that is fuzzy at best.

By 'reasoning' Denialists seem to mostly mean our ability to grasp evaluative facts on the basis of evidence (*Ibid.* p. 290). Our reflective intuitions are a part of that evidence, but other scientific theories, like evolutionary theory or experimental psychology, also are (pp. 294-296). In practice, while Denialists often *claim* that they are arguing broadly against 'common intuitions' in favor of 'rationalism', when you examine their actual arguments what they are really saying is that *some* of the reflective intuitions we derive our moral theories are unreliable while others (such as the all-things considered *intuition* that utility should be maximized) are more reliable, as are reasons derived from scientific evidence like experimental psychology. But even granting that does not allow escape from the Reidian argument.

Even if Denialists *can* differentiate between intuitions and reasoning, they still have to a) provide some argument for why the latter is more reliable than the former, *and*, to get the conclusions they actually want, b) provide an argument that our more reliable sources of moral knowledge favor consequentialism rather than nonconsequentialism. The first challenge is exactly where the Mulgan/Reid approach targets – many possible criticisms that cast doubt on the reliability of our reflective intuitions also apply just as well to our reasoning ability, moral or otherwise. But even if they answer that, it would still leave the final hurdle, which is I think the

most dubious, and where they overreach the most. I will go over their arguments in more detail below, but to summarize I argue that while you can make good arguments that *certain* intuitions are unreliable, or rescue certain moral judgements from debunking arguments, you cannot do so with the generality required to make the claim that 'intuitions are less reliable than reasoning'. Furthermore the reliable and unreliable moral judgements don't map respectively onto consequentialist/nonconsequentialist conclusions in any kind of systematic fashion.

The various attempts to separate our intuitions as being particularly unreliable when compared to our other sources of moral knowledge can be divided into three broad categories, though they are not mutually exclusive, These are a) attacking our intuitions on the basis of their evolutionary origin, b) trying to draw a distinction using empirical moral psychology and c) Unger's strategy, which is to divide our general intuitions from our intuitions about specific cases. I will go over each in turn, but the general structure of my counterargument is similar in all cases.

Evolutionary debunking arguments

Singer's argument against our intuition is a form of evolutionary debunking argument. Evolutionary debunking arguments cast doubt on the reliability of various human capabilities on the grounds that those capabilities are the result of processes that select for *reproductive fitness*, rather than reliability or correctness or whatever it is that is being cast doubt on. This is meant to undermine our belief in the reliability or accuracy of the capability, since it gives us reason to believe that our capabilities are only reliable or accurate to the extent that that contributes to fitness. Even worse, it gives us reason to believe they are unreliable if that unreliability contributes to fitness: to give an example I will return to, our moral intuitions are likely to stray towards selfishness.

But evolutionary debunking arguments need to be a little more complicated than that if they are to actually work. The mere fact that a capability is the result of a process that selects for something other than accuracy need not mean that it is inaccurate, because oftentimes a capability *will* give more reproductive fitness the more accurate it is. The example I would give is the same one that Zachary Ernst gives in an article discussing anti-intuitionist consequentialist arguments (Ernst, 2007): our perceptual facilities. In fairness, it is not nearly as plausible that our moral instincts are adaptive *in virtue* of generating reliable moral judgements the way our perceptual faculties are generally adaptive in virtue of generating reliable perceptual judgements[4]. It is, however, possible that they track something *coextensive* with moral properties which gives them a reason to be accurate to some extent. To give a clarifying example, we did not develop the capability to identify and roughly measure the wellbeing of others because doing so would be useful for moral theorizing, but rather for other reasons. That does not change the fact that our intuitions on this matter are useful for moral theorizing. It is not actually necessary that our capacity to be moral is the direct result of evolution at all: we could have developed a number of other capabilities (compassion, empathy, reasoning) that are adaptive for *other* reasons and thus generally reliable, but which happen to allow us to make accurate moral judgements despite not having evolved for that purpose.

Singer himself uses these kind of arguments to defend our moral reasoning abilities. In a recent paper he wrote in collaboration with Katarzyna de Lazari-Radek (Singer & de Lazari-Radek, 2017), Singer talks about how to combat Sharon Street's "Darwinian Dilemma" (Street, 2006), which is also an evolutionary debunking argument but in the realm of metaethics. Street argues that moral realists face a dilemma: either our evaluative attitudes evolved due to processes

[4] At least, about everyday objects.

that are not truth-tracking – and are therefore unreliable – or evolutionally forces selected for evaluative attitudes that pointed to the truth. In Street's view, that latter option is scientifically untenable, because it simply much more plausible that our evaluative attitudes evolved in ways that are evolutionary advantageous rather than that they tracked the truth.

Singer and de Lazari-Radek accept the first horn of Street's dilemma with respect to most of our moral intuitions (e.g. the intuition that distance is a morally relevant consideration), agreeing that they are unreliable, but try to rebut the second horn for other moral judgements that they argue arise from our more basic ability to grasp evaluative facts (e.g. the judgement that we should try to maximize the utility of all sentient beings). Arguing that our ability to grasp moral truths is an aspect of our ability to reason, and that there is of course no mystery or difficulty in explaining why our ability to discern basic evaluative facts (e.g. this food is better to eat than that) might have evolved to be generally truth-tracking, Singer and de Lazari-Radek reject that taking the second horn of the dilemma is unviable or scientifically untenable.

This is where the vagueness of throwing around terms like 'reasoning' and 'intuition' without significant clarification starts to become particularly unhelpful, but specific examples may help. For instance, it is likely we developed the ability to feel compassion because it caused us to aid those close to us and, because those close to us are usually related to us[5], this is evolutionarily advantageous thanks to kin selection. But that doesn't mean that we *only* feel compassion for those related to us, because evolutionary forces have limits and simply could not instill so specific a condition on our compassion. Because several of our intuitions evolved to reliably track those actions that would improve the wellbeing of kin – which, for most of human

[5] At least, in the circumstances under which the capacity to feel compassion evolved

history, meant the majority of people you would interact with – those same intuitions *also* will reliably track those actions that improve the wellbeing of people in general, because ultimately people aren't different enough for them not to. Now, they will likely tend to be more *biased* in the favor of in-groups and kin than they would be if they had evolved to simply be truth tracking, but that doesn't make them useless for moral theorizing, it just means that they have a particular flaw. They can still be reliable guides to moral theorizing provided that flaw can be corrected for – and while they don't go to the extremes that consequentialists do, even the most ardent intuitionist generally agrees that the correct moral theory will be *somewhat* more impartial than our basest intuition, i.e. that our intuitions need to undergo a process of reflection before we can rely on them. On the flipside of that argument, it is possible if not likely that our ability to reason about basic evaluative facts *also* has flaws and biases that need to be corrected for when it is applied to moral judgements, as a result of being used to do something it is not quite designed to do.

I am not arguing for any of these theories of how our moral capacities evolved on empirical grounds (evolutionary psychology is a field in which it is notoriously difficult to make good empirical arguments), but trying to point out that there are in fact *many* ways in which our moral intuitions might have evolved to be reliable in at least some ways (while being unreliable in some other ways) despite evolutionary pressures not driving them to be truth tracking. This isn't even the limit of the possibilities – it is entirely possible that our ability to be moral is in fact *nonadaptive* (in the sense that it can cause us to do things that reduce our reproductive fitness) and would be selected against, save that it is a natural consequence of the *combination* of several other of our capabilities (reasoning, empathy, compassion) that each are of the type that they must be reliable if they are to increase reproductive fitness and are each *individually* too

valuable to lose. In this kind of scenario, evolution cannot get rid of our ability to find moral truths because it would require us to lose adaptively useful capabilities, just like it cannot get rid of the blind spot in the center of our vision because it would require us to lose our eyes and re-evolve them from scratch so that they are wired the right way around.

This might seem a remarkable coincidence, that a combination of capabilities not individually evolved to do so can grasp onto moral truths, but if you are a moral naturalist it doesn't have to be. Many of our evolved capabilities *have* to be sensitive to natural facts if they are to be adaptively useful, but that means they are also capable of being sensitive to natural facts in general. In this sense, we are capable of developing accurate moral theories similarly to how we are able to develop accurate physical theories – as an extension of our abilities to discover natural facts about the world, rather than via an ability developed specifically for that purpose. A moral non-naturalist has a much harder time explaining this, but this is not anything new for them. Non-naturalist moral realism has always had an epistemological problem, with one of its greatest challenges being explaining how we come to have accurate moral knowledge. The evolutionary debunking argument does not make this challenge noticeably harder than it already was.

But this is a rebutal that applies to both horns of the Darwinian dillema, not only the second one. Just as our reasoning may be able to grasp moral truths despite not having been specifically evolved for doing so, so the same may be true of many other human capabilities, including our intuitions. This is the flipside of the Reidian argument – any argument that can recover critical consequentialist claims like utility maximization from the Evolutionary Debunking Argument can *also* rescue critical nonconsequentialist claims just as easily. For example, I would argue that the same arguments Singer and de Lazari-Radek use to defend the

Golden Rule and utility maximisation (Singer & de Lazari-Radek, 2017, pp. 288-290) against EDAs can also save, for example, a constraint against doing harm except in self-defence. I agree with Singer's and de Lazari-Radek's arguments and have never found Street's Darwinian Dilemma to be very compelling for exactly these reasons. But this undermines Singer's own attempts (Singer, 2005) to deploy the EDA to defend consequentialism from unintuitive conclusions by casting doubt on the intutions it clashes with, since many (though not all) of those intuitions can be rescued by the very same arguments he uses here to rescue utility maximisation from Street's Darwinian Dilemma.

And of course if you are not either kind of moral realist there is another way to avoid the Evolutionary Debunking Argument, and that is to take Street's side of the issue. For some types of metaethical theories, which Street somewhat inaccurately[6] labels 'anti-realist' but which are perhaps more usefully labeled 'perspectivalist', our evaluative attitudes are in some way the source or determinant of accurate moral beliefs, and so our intuitions are naturally in some way going to track the latter. That doesn't mean that a perspectivalist isn't going to judge some sources of moral knowledge as being better than others, for there are still various reasons why they might do so. But it does mean that an overall evolutionary debunking argument loses most if not all of its force, since there is now a simple explanation for why our moral attitudes might track the moral truth: because said truth is in part a function of those attitudes.

In the final analysis, there may well be reasons to doubt certain of our intuitions on evolutionary grounds – it is likely, for instance, that our evolutionary history causes us to

[6] It's inaccurate because many of the views that Street lumps under the label – constructivism, for example – tend to be held as realist by those who hold them. At one point, she even claims that a particular type of value *realism* escapes the dilemma but is anti-realist by the standard she is using (because it has those values be in some way mind-dependent) (Street, 2006, p. 136) which is further reason to think it is the wrong word.

prioritize our group (family, tribe, nation, etc.) more than is actually morally justified, as Singer and de Lazari-Radek note (Singer & de Lazari-Radek, 2017, pp. 291-292). But an overly broad evolutionary debunking argument is not only self-defeating in the Reidian sense, there are many ways in which a faculty might have evolved to be accurate even despite evolution not selecting for that directly. *Specific* evolutionary debunking arguments might work, casting doubt on particular moral intuitions (e.g., tribalism, incest taboos) due to their evolutionary history, but that does not suffice to cast doubt on all our non-reasoning based sources of moral knowledge, even if one accepts the already somewhat dubious claim that it is at all simple to draw a sharp line between our 'reasoning' abilities and the other ways we might come to moral knowledge even on evolutionary grounds. But the main point I want to emphasize here is that there is a large difference between thinking that evolutionary debunking arguments give us good reason to think that our moral attitudes have certain biases or flaws, and thinking we should discard our intuitions altogether.

Arguments from empirical psychology

In the face of the counterargument that any doubt cast on our intuitions due to their evolutionary origin would be equally easily cast on our reasoning, and any *defense* of our reasoning might as well be applied to our intuitions, anti-intuitionists like Singer must provide additional arguments both for distinguishing the two and explaining why intuitions are *especially* problematic. One such set of arguments relies on empirical research in psychology; such as those by Joshua Greene (Greene et al. 2004) (2008) (2009) (2010). Greene et al. examine the neurological activity of subjects who are making moral decisions using fMRI scans and other methods, such as measuring reaction times. As a result of his research he advocates for a 'dual-process' or 'multi-system' theory of moral psychology.

The basic idea is that we have two different subsystems involved in making moral judgements, one of which is more emotional while the other is more cognitive. In the example of trolley problems that Greene uses for the purposes of his research, Greene demonstrates over the course of multiple experiments that people who make consequentialist judgements (e.g. willing to push a Fat Man onto the train tracks to save five people) use the more cognitive system while those who make more deontological judgements (not pushing the Fat Man) are making more emotional judgements. Greene wants to argue from this that deontological theories such as Kantianism, because they accommodate the more emotional moral responses and because those responses are particularly suspect to the evolutionary debunking argument, are more likely to rely on non-truth tracking intuitions than consequentialist moral theories that rely more on the cognitive subsystem. Therefore, we ought to reject Kantianism and other deontological moral theories (2008).

For the sake of argument, let us accept Greene's empirical claims. There have been some scientific criticisms of some of his experiments, which I do not feel qualified to judge for myself, but overall I suspect we have good reason to believe in something *like* the dual-process theory. I have personally long felt that a multi-system theory is the most plausible theory of moral psychology, as it is likely we evolved different kinds of moral intuitions in different circumstances and for different reasons. Many moral psychologists have advanced theories along these lines, most famously Jonathan Haidt (2012). Let us accept that in the Fat Man case, the deontological judgement is the more emotional and the consequentialist judgement the more cognitive. It is still a long way to get from there to a broad rejection of deontological moral theories.

Greene's experiments focus heavily on trolley problems, and investigate a particular sort of emotional response we have that can lead us to making moral judgements – the aversion to the use of personal force. Most philosophers agree that whether you kill someone up close and personal or by pressing a button from a distant bunker is not a *morally* relevant difference, but our emotional responses will nonetheless differ, and this will influence our moral judgements. Greene uses examples like these to argue that our emotional responses too often cause us to consider moral irrelevant things, and thus we should discount the more emotionally driven moral subsystem in favor of the more cognitive one.

But notice something about that example: most philosophers, even most consequentialists, would say that people, as a general rule, are *more* comfortable with remote killing (via the use of drone warfare, to give a topical example) than they ought to be, not less. In this case, the lack of emotional response is doing more to lead us astray than the emotional response is. Greene might respond that he is not denying that the emotional response to killing is an important motivating factor for preventing people from *acting* immorally, but when we do moral *theorizing* we should nonetheless discount our emotionally derived intuitions. But this counterargument still casts some doubt on that assertion – sometimes our emotional responses can nonetheless lead us to moral action better than cognitive reasoning can, especially since cognitive responses can be misled due to limited or ill-considered information. I will return to this point in a moment.

But the main objection to Greene is that he makes a broader argument than he is licensed to from his experiments. Greene wants to argue that deontological theories are too accommodating to our emotional moral subsystem, but all he's really shown is that in *trolley problems* the nonconsequentialist response is more emotionally driven. Given the sheer breadth

of moral theories that classify as 'deontological', it is absurd to argue that all nonconsequentialist theories rely on suspect moral subsystems and so we should be consequentialists. Kant, after all, believed that his theory was the result of pure reason and had his own share of nonintuitive conclusions. Greene oversteps when he calls all deontological moral theories suspect on the basis of a single case where they are more emotionally based.[7]

There is a parallel here with the other forms of the evolutionary debunking argument. We may very well have good reason to suspect that some of our moral intuitions arise from emotional reactions (disgust, revulsion, etc.) that are unlikely to be truth tracking, just as we likely have good reason to think that our intuitions have certain biases as a result of our evolutionary history. But our reasoning capabilities are also the result of evolution and have their own biases, and it is difficult to conceive of an argument that could explain how our reasoning abilities can be truth tracking that also cannot be applied to rescue some of our other sources of moral knowledge as well. To give an example, we are much better at reasoning involving social situations than with highly abstract reasoning, even for problems which are logically identical, and it is likely because our reasoning specifically evolved for the former case and can only incidentally do purely abstract reasoning (Cosmides & Tooby, 1992). But this means that, as in the drone case, we are actually more likely to go astray the more we abstract away our emotions, and we may be better off not doing so.

A wholesale anti-intuition argument is both implausible and self-defeating, and Singer and Greene and other consequentialists are aware of this. That is why they try to provide reasons to doubt *some* of our intuitions – the ones that lead to nonconsequentialism – while accepting

[7] For a more in-depth version of this argument, I recommend Meyers' paper "Brains, trolleys, and intuitions: Defending deontology from the Greene/Singer argument" (Meyers, 2014)

others. But in this, they overreach: neither their arguments nor the empirical data as yet give us blanket reasons to doubt all nonconsequentialist intuitions, only reasons to doubt some particular kinds of intuitions. Now, I do agree that there are certain subsets of our intuitions that are particularly unreliable and, as a result, *certain* arguments against consequentialism lose much of their force. But it is a long leap from there to the broader argument trying to undermine *all* nonconsequentialist intuitions.

In particular, I do not think the sets of our intuitions that are reliable or unreliable map on in any way neatly to those sets that support consequentialism or nonconsequentialism respectively, nor do I think that Singer or Greene have given us any good reason to think so. Part of my purpose in this project of mine is to try and identify what intuitions we can trust and which we need to discard, but in this I am no different from any moral theory, all of which must clash with our intuitions at some point[8]. But I do not think one can argue on evolutionary or psychological grounds that our consequentialist intuitions are generally more reliable than our nonconsequentialist ones.

Unger's argument from deeply held moral principles

Peter Unger (1996) follows a different line of argument: that our intuitions about particular cases are more unreliable than our intuitions about general moral truths. The latter, he says, reflect our 'primary' – most basic – moral values, while the former reflect our secondary or derived moral values. Whenever these two values conflict, we should favor the primary over the secondary. Unger uses this argument to undermine objections to demanding moral theories that

[8] Even a purely intuitionist theory – one that simply claims that what is right to do is what we instinctively think is the right thing to do - will do so, since our intuitions sometimes clash with each other.

require one to give substantially more to charity than most people do, though he avoids nailing himself down to any particular moral theory.

But Unger is, once again, making claims that don't follow from that distinction alone. Even if he can distinguish our primary moral values from our secondary moral values, it not at all clear that he is licensed to claim that the latter are less *important* simply because they are less *general*. Garrett Cullity notes: "to most people, it is about as obvious that there is a moral difference between our relations to a child drowning in front of us and a child starving in another country as it is that failing to save a drowning child is wrong" (1994). In other words, the two intuitions may be of different levels of generality, but they are of equal strength: Unger cannot discard one and not the other on the basis of generality alone. And as discussed above, many of the other ways one might distinguish between them are suspect for very similar reasons.

Unger may well be right that the kinds of judgements involved in particular cases are influenced by other considerations than the basic values that he thinks they ought to solely be considering. But it does not follow from that that he is right that all of those considerations ought to be dismissed (Ernst, 2007). Though some might well! Once again, I do actually think that Unger correctly identifies in many of the cases he considers factors affecting our judgments that are genuinely extraneous. Just as before, my criticism is about him overgeneralizing from those cases. And also just as before, his arguments for why our primary values are more important than our secondary ones are subject to just the same Reidian counter-argument that they source of the two kinds of values is too similar for us to be able to criticize one sort without criticizing the other, though we may have good reasons to doubt some particular values of *either* kind.

The consequentialist case for intuitions

In summary, I don't think that skepticism about intuitions can be used to argue for consequentialism. I agree that we have good reasons to think that some of our sources of moral knowledge are more reliable and others less so. But what we don't have good reason to think is that this division maps neatly onto either the reasoning/intuition distinction *or* the consequentialist/nonconsequentialist distinction. Indeed, I think that one can even make a *consequentialist* argument that as long as our intuitions are on roughly the right track than our nonconsequentialist intuitions may sometimes be a more reliable guide to action than pure reasoning.

The basic outline of this argument is that our intuitions are the result of an evolutionary process that has 'priced in' all of the long-term and invisible consequences of a course of action that any realistically doable calculus might miss. I am far from the first person to argue that consequentialists should generally follow common-sense moral rules in the majority of cases: it has been a staple of consequentialist thought from the beginning[9]. So far so simple, but the argument can be extended to more difficult cases as well. Imagine circumstances like the common thought-experiment of Transplant, or similar cases where consequentialism seems to argue for courses of action we normally think of as immoral. Another staple consequentialist argument is that in realistic circumstances consequentialism does not actually recommend the immoral course of action – it only does so in thought experiments where all the possible outside

[9] See, e.g., (Bentham J. , 1789, pp. Chap IV, Sec. VI) (Mill, 1861, pp. Chap II, Par. 19)

considerations and long-term consequences are removed, circumstances so unrealistic that our intuitions are no longer a reliable guide[10].

This sort of argument is often presented as an *anti*-intuitionist argument, a reason to doubt our intuitive responses to the Transplant case. But it has a flipside: it means that the consequentialist is arguing that in *realistic* circumstances, our intuitions *are* correctly labeling an action as immoral that does actually have negative consequences. A too shallow consequentialist calculus, on the other hand, would permit the action – this is after all what gives cases like Transplant argumentative force in the first place. Thus, there is a *consequentialist* argument that it is better to go with our common-sense intuitions then to do a shallow calculus and – since the kind of in-depth calculations that would accurately take into account *all* consequences are often impractical or impossible for reasons of time, lack of information, or uncertainty – that in turn means that sometimes consequentialists recommend going with your intuitions over your reasoning. So even from a purely consequentialist point of view, it's not nearly as simple as saying that reasoning is always better than intuition.

This is why I think that a moral theory must *interrogate* our intuitions, but not abandon them. Our intuitions are the result of imperfect processes and, thus, prone to failure. But our reasoning is also imperfect and prone to failure, and without our intuitions to guide us we would be entirely lost. A moral theory gives us a means to reflect on our intuitions and at the end of the process we may have to adjust them or even discard them entirely; but it ignores them at its own peril. This then, is also a summary of the rest of this project: an interrogation of our intuitions through what I think is a useful consequentialist lens that helps us separate those of our intuitions

[10] See, for instance, (Kagan S. , 1998, p. 77)

that are on the right track from those that are more likely to be mistaken. Our intuitions might point us to the right direction but, to continue a by now perhaps mangled metaphor, they won't get us all the way to the end. It is my hope that this project of mine, second order consequentialism, will lead us further on this track. At the very least, my hope is that reexamining our intuitions through this light will give us some insight into how reliable they truly are, and which ones we should keep or discard.

Conclusion

To sum up, I think there might be ways to distinguish between 'intuitions' and 'reasoning'. I believe that we have good reasons to think that some of our sources of moral knowledge are more reliable than others, both from empirical research and simply because it would be an extraordinary coincidence if all our sources of moral knowledge were equally reliable. But I do not think it generally plausible that our nonconsequentialist intuitions are less reliable than our consequentialist ones, due to all the arguments outlined above[11]. Nor do I think we can generally say that our reasoning is more reliable than our intuitions.

Thus, while I do aim to use a form of consequentialism to interrogate our intuitions, I don't think one can use anti-intuitionist arguments to argue for consequentialism itself. This chapter should have made clear why I believe that line of argument is untenable, even self-defeating. More than once, I come to the conclusion that our intuitions supporting what ordinary or *first*-order consequentialism might say about a particular case might be the more unreliable rather than the reverse, as I sketched above. My reasons for nonetheless adopting a strategy that

[11] Though I do think that when it comes to the question of demandingness, we have good reason to cast especial doubt on our nonconsequentialist intuitions. But that is due to *specific features* of that case. I will return to this argument in more detail in Chapter 6.

is overall consequentialist, and which analyses and criticizes our intuitions on those terms, will be the topic of the next chapter.

Chapter 3: Why Consequentialism?

Introduction

Consequentialism is a set of theories that are well known and long discussed, and therefore so are their problems and possible solutions to the same. In this chapter, I will go over what I think are the most compelling reasons one might be attracted to consequentialist theories, and also briefly go over some of the most major problems. I will argue that consequentialism is a theory worth exploring despite its flaws, but that many attempts to fix the problems with the theory – satisficing consequentialism, for example – end up altering it in ways that undermine the initial reasons one might have for adopting consequentialism in the first place. This is because they tend to operate by adding modifiers or introducing new factors to a consequentialist base, complicating a theory that has a foundational simplicity as one of its main draws. A true defense of consequentialism, in my view, must provide a way of answering these objections that nonetheless does not rely on introducing piecemeal rules or adjustments. In this chapter I will explain why I think that one of consequentialism's greatest advantages is its simplicity and unity of explanation and that therefore we need a systemic approach to fixing its flaws rather than a series of independent adjustments. What I think such a systemic approach might look like is the subject of the remainder of this work.

Consequentialism

My personally preferred definition of consequentialism is that a theory is consequentialist if it places primacy on making the world a better place over all else. In its most straightforward and familiar form, this generally translates to saying that the right thing to do is that which makes the world best off. Most people (with a very small number of exceptions) of course think that

making the world a better place is a good thing to do, and even a moral imperative, but consequentialists believe it is what doing good primarily *consists of*. The contrapositive makes it easier to see the fault lines: a consequentialist believes that making the world a better place is never the wrong thing to do. A nonconsequentialist, on the other hand, holds that there are actions that are wrong even if they make the world a better place (the most straightforward example being murder), or – more rarely – right even if they make the world a worse one.

This is a very simple statement of consequentialism, and often does not precisely apply to more sophisticated variants, yet in some ways it is the most compelling argument for it. It is very appealing to have a moral theory that says it is never right to make things worse, and always good to make things better. This doesn't feel like a statement that itself needs justification: rather, it is the nonconsequentialist, who says that sometimes the right thing to do is to make the world *worse* off, that seems to require a justification. That said, this is not as difficult a justification as it might seem at first glance: we already mentioned the case of murder. Even if one holds that a world is the worse off for having a murder in it, which blocks off most truly intuitively offensive scenarios where consequentialists might recommend murder, it still seems that consequentialism would allow for murder to prevent five other murders[12]. If one rejects that, one also must reject consequentialism. Nonetheless, this quite simple core of consequentialism is something that I think many consequentialist theories, even as they become ever more sophisticated to account for things like murder, try to preserve. I, too, wish to preserve it: my theory is quite different, in the end, from standard maximizing act consequentialism, but to some

[12] One can possibly get around *this* by having goodness be agent-relative, which I will discuss later in the chapter, but in general one can engineer scenarios where consequentialist theories recommend courses of action we might wish to be blocked off by hard constraints.

extent I still want to say that an action is right only if it (somehow, eventually, in the long run) makes the world a better place.

Another positive argument for consequentialism[13] is that we all care about consequences to at least some extent. Most people think we have some moral imperative to better the world; even arch-deontologist Kant held it to be an imperfect duty. There are some people who are exceptions to this or at least sometimes *seem* like they might disagree with this, and we will discuss some of them in more detail later on in this chapter, but the very basic claim that making the world better is generally the right thing to do is mostly accepted. If then, the argument goes, consequentialists can show that this simple principle is *sufficient* – if we can get everything we want out of our moral system from consequentialism alone – then they do not need *further* arguments in favor of their theory, since the core of it is something already agreed to. This is of course a far from trivial exercise, but like the previous argument it largely puts the burden of proof on the nonconsequentialist side.

Arguments that rely on shifting the burden of proof, however, can often result in a futile tug of war. Even granted that most people agree that consequences matter, if one starts with a theoretical foundation that simply produces the duty to better the world as one among several duties (as a Kantian might), they may say that it is the consequentialist that must justify this duty being somehow special and distinct from the others. If constructing a public-reason-like argument that 'bettering the world' is notable because it is a shared reason accepted by all moral theorists, many other nonconsequentialist duties or rights are similarly widely accepted. The imposition against doing harm, for instance, seems at least as basic. Let us say that the

[13] e.g. Shelly Kagan (1998)

consequentialist *can* produce a theory that gets us all we want out of morality. This surely means that they must accommodate intuitions like the prohibition against murder even when it might better the world. Similarly, a deontological theory that also gives us all we want out of morality must also accommodate the consequentialist intuition. If we assume for a moment that both theories exist, the mere fact that we all care about consequences does not seem to give us reason to adopt the hypothetical complete consequentialist theory *over* the hypothetical complete deontological theory.

But the Kagan argument does, I think, give us a reason to at least investigate the possibility of consequentialism. That is, it tells us that it is worth seeing if we can have a theory that from the idea of bettering the world alone produces what we want out of a moral theory. Fundamentally, what we do as ethicists is try to start from principles we can all agree to and produce from there a theory that can guide us in less certain situations, consequentialism is merely one such starting point. But as for a reason why we should take this as our starting point as opposed to something else, there I think the best argument is in fact the innately appealing notion that bettering the world cannot be wrong.

Simplicity, aggregation, and unity of explanation

Another argument that might be made in favor of consequentialism it that it has simplicity. Not simplicity in the sense of being unsophisticated, but rather in the sense of being elegant. It gets more accomplished with fewer premises, it has elegance in the way a mathematical theorem that proves the unobvious from the simple in a few clear steps has elegance. Like the shifting of the burden of proof, though, arguments from Occam's Razor tend to go back and forth based on the angle one is approaching the problem from. There is surely *a* sense in which consequentialism, especially maximizing act consequentialism, is simple. It has

only one primary principle: to make the world the best it can be, to do that act that has the best consequences. But it can be very legitimately argued that this simple principle is, in another sense, too complicated. When it comes to practical decision making, it is actually *harder* to apply this principle in many real world situations than it is to adhere to a set of simple rules, even possibly quite a long set of rules. A deontologist, who advocates adhering to a system of rules, might well argue that theirs is the simpler moral system to actually implement, and I think there is quite a good argument to be made in their favor.

It is in defense against this claim that I think the Kagan argument is best deployed, rather than as a positive argument in favor of consequentialism. Yes, transforming a principle like "make the worst the best it can be" into actual real world decisions is a very complicated and not at all simple enterprise, but it is also not an enterprise only the consequentialist is saddled with. If you care *at all* about consequences, about making the world a better place, then you too must grapple with questions such as "better is what way?", "how do we measure that?", "how do we translate that into guides for decision making?". Indeed, the nonconsequentialist has arguably a harder task here, because they must additionally balance these questions with all their nonconsequentialist concerns, whatever those might be. The only way to avoid this difficult problem of turning the consequentialist principle into a guide to practical decision making is to abandon the value of consequences altogether, to argue that bettering the world is simply not among the things we ought to do. And not many people are willing to go that far.

That said, there are some philosophers who at least *seem* to say something along those lines. However I believe that when one examines their arguments they are not actually arguing against the idea of assigning importance to consequences at all, but rather against certain ways of weighing those consequences. Rawls' separateness of persons argument (Rawls, 1971, p. 37), for

example, specifically targets utilitarianism, and a consequentialist theory that takes distribution into account evades it. And of course Rawls himself is very concerned with making people better off. Even the most extreme arguments against utilitarianism don't go so far as to say that consequences are not part of the equation. The philosopher Taurek infamously argued (Taurek, 1977) that numbers do not matter, that the goods of individuals cannot add up to the goods of a 'overall' good in any meaningful sense. But as much as this argument would dismiss traditional consequentialism as incoherent, it still accepts that there *exist* goods of individuals. Taurek merely must now produce a different set of decision making criteria we must use to evaluate how to best serve those goods. Perhaps, as he argues, if given a choice to save one or many we should flip a coin so that everyone has an equal 50% chance of survival. If he is right about that, it has massive implications for our practical ethics, our day to day decision making, our government policies, everything about our moral theories. But that would be true for *everyone*, not just consequentialists.

A very similar thing can be said for the idea of incommensurable values: values that have no common standard of measurement and so therefore seem difficult to directly compare with each other, at least in many circumstances. Many of us, even most consequentialists, are pluralist about values. Arguably even a utilitarian, at least one who acknowledges dimensions and qualitative aspects to pleasure such as John Stuart Mill (Mill, 1861), must contend with the problem of weighing different *types* of pleasures against each other, which means many of the same issues (Chang, 1997). Obviously if some values are incommensurable with others it presents a problem for consequentialism, but it is even more difficult to conceive of a moral theory that does not *ever* have to weigh values against each other than it is to conceive of one that does not care about consequences at all. And indeed, unlike with the problem of aggregation

which often arises in criticisms against simple utilitarianism, incommensurability is usually recognized as a general problem that afflicts most moral theories, and discussion tends to focus on ways we might make decisions even if our values are incommensurable[14].

One incommensurability claim that does seem to specifically target common forms of consequentialism is the claim that human lives are special and cannot be weighed against other kinds of benefits. Like the challenge from aggregation, this is often presented as a claim that utilitarianism and other similar forms of consequentialism do not properly respect human dignity. On this matter I tend to side with Alistair Norcross (1998) whose arguments can be summarized quickly as: we regard actions that increase the risk of lots of people dying to be as heinous as those that will certainly kill one person, yet we often weigh the risks of lots of people dying against other goods. I will discuss this argument in more detail in chapter 12, for now I want to emphasize less the part of the argument that claims we are allowed to put lives on a balance with other things, and more the part that notes that we *do* put lives on a balance with all the goods, all the time, when we are debating government policy or safety precautions or any number of other things. The point, again, is that if we *are* truly incorrect in doing so it is not some sort of special problem that only utilitarians have to worry about, it is a problem with *any* kind of moral theory that wants to have something to say about practical problems such as government policy – which surely is any moral theory worth considering.

In the end, even if we cannot aggregate multiple goods and different goods are incommensurable, this does not mean that consequentialism cannot claim *relative* simplicity. Making a consequentialist theory that produces a practical guide to decision making in the face

[14] see e.g. (Parfit, Reasons and Persons, 1987, p. 431) (Taylor, 1982) (Stocker, 1997)

of these difficulties might seem an almost insurmountable task, but the nonconsequentialist would face all the same issues with doing so – and also have to account for their nonconsequentialist concerns as well. In other words, problems of aggregations and commensurability are not problems for consequentialism, broadly construed (they are certainly problems for many of the most common forms of consequentialism), they are problems for any moral theory. They do not succeed in weakening consequentialism's claim to simplicity. What they would do, of course, is make it far less plausible that we *could* get everything we want from a moral theory out of consequentialism alone. But if we assume for the moment that we could, the appeal to simplicity remains as an argument in favor of consequentialism, a reason to adopt it over other theories, even ones that are equally successful in terms of getting what we want out of morality.

A perhaps better way of putting the notion of 'elegance', I think, is describing it rather as *unity of explanation*. A consequentialist theory might need a complicated apparatus in order to translate it into a decision making heuristic, but at its core is a simple straightforward principle: make the world a better place. The advantage here is more than just simplicity: just as a value pluralist must deal with the problem of how to weigh different values against each other, and with the specter of incommensurability and incomparability making things even more difficult in that regard, so too must the *principle* pluralist have to deal with the tricky issue of what to do when those principles collide. If – and granted that it is a rather large if – this one principle alone can get us everything we want out of morality, then that would mean there is an entire class of tricky moral problems and complex issues that we might be able to entirely avoid. Entirely separate from the bare appeal of simplicity as an end in itself, this possible benefit seems to me to make it worth seeing if we can get a consequentialist theory to work.

The problems with consequentialism and piecemeal solutions

Thus far I have been arguing for the attractiveness of consequentialism providing it can give us everything we want out of morality. But of course most nonconsequentialists are so because consequentialism *doesn't* give them what they want out of a moral theory. It is too permissive, it's too demanding, it's too hard to aggregate and measure consequences, and so on. There are a myriad of problems with consequentialism, and in later chapters I will discuss the most prominent ones in detail and go over my own solutions. For now, I wish to bring up what I see as a general problem with many proposed solutions, and why I see my overall project as an attempt to address this problem.

In response to the criticisms of consequentialism, consequentialists tend to take one of two options. The first is the Singer/Unger route we discussed in the previous chapter, rejecting the criticism and maintaining simple consequentialism. The second option is to accept the critique and modify their consequentialist theory accordingly, adapting it or sophisticating it in response to the criticism. Often this manifests as altering ones theory of the good – including equality as a good in itself in response to a Rawlsian style critique, for example – or by adding or altering consequentialism away from simple maximization. These alterations may sometimes have their own problems, which I will discuss in more detail in later chapters, but I am interested in this chapter not in their individual issues but the problem that arises when many such alterations are added to consequentialism.

If one adapts their consequentialist theory to respond to criticisms or to conform more closely to our intuitions, the result is a changed theory, and generally a more complicated or

sophisticated one. For instance, in response to the charge that consequentialism is too demanding one may move to satisficing forms of consequentialism which allow for actions that have less than the maximally best consequences provided that their consequences are overall good enough. But now to the already difficult problems of consequentialism, such as "how do we measure consequences" and "how do we work out what actions have the best consequences" we have added the additional problems of "how much good is 'enough' good" and "why that much and not some other 'enough'". This is already going to make things more difficult by itself, but demandingness is far from the only place where consequentialism clashes with our intuitions. And while each individual modification may be simple, if we continue to add modifications to solve other problems on a piecemeal basis the result will be unrecognizable compared to the starting theory and risks losing the features that attracted us to it in the first place.

This issue that answering criticisms might require consequentialism to become too complex is a common problem in philosophy that arises in many different fields. Philosophers, as a rule, are pretty clever people. They can often work out ways to make almost any theory *work*. By massaging definitions, adding stipulations, explaining away seeming contradictions, one can take almost any theory and get something at the end of it that is immune to the most common critiques against it. Just like how a good lawyer can construct *a* legal argument in favor of almost anything, a good philosopher can contort a theory to evade almost any critique. The real challenge is often not that part, but rather doing so without losing the initial advantages that would have drawn one to the theory in the first place[15].

[15] The other challenge, alas, is catching ourselves when we are doing this.

In the case of consequentialism, the most obvious challenge is that if one of the draws of consequentialism is the relative simplicity and the unity of explanation, every additional stipulation or condition is a threat to that. But even setting that aside, the more complex and sophisticated a consequentialist theory gets, the more it is in danger of losing the other main draw of consequentialism, the clear and strong connection between doing the right thing and making the world a better place. Now it might still be the case that a sophisticated and indirect consequentialism might have advantages over nonconsequentialist theories in some cases (one example we will discuss later being the case of thresholds) but if one has lost that clear connection, one has also lost much of the motivation to be a consequentialist. Especially if one can then get what they want out of morality more straightforwardly by some kind of nonconsequentialism, at which point consequentialism has also lost any claim to relative simplicity.

At this point one might be tempted to instead stick with simple consequentialism regardless of its problems, and work to challenge and minimize the critiques rather than adjusting consequentialism to fit. But as I discussed in the previous chapter, I think that project has some success in places but does not work as a general solution. Unless one discards *all* our common intuitions as unreliable, and as I said in the previous chapter I ultimately think we are not justified in doing so, there are going to be cases where simple consequentialism must undergo alteration or face irreparable conflict with cherished intuitions.

My thesis: moving to a different level

The central thesis of this project is that there is a way to retain a large part of the traditional advantages of consequentialism while still resulting in a moral theory that is responsive to the common critiques against consequentialism. This is accomplished by applying

consequentialist reasoning not to actions directly but rather to the ethical theories we adopt to guide our actions. This idea is most well known in the form of rule consequentialism, but as I am not wedded to the idea of rules I call the more general version **second order consequentialism**. The idea here is to retain the relative simplicity and especially the unity of explanation. Even if it is not on the level of action, this still allows us to avoid some of the difficulties of pluralism about core principles, because we have our second order consequentialist theory that tells us what to do when those principles conflict.

But moving away from applying consequentialism directly to actions comes with risks of its own, namely once again weakening that strong connection between the right thing and making the world better. My position is not that we should do that which makes the world best off, but that we should adopt the moral system which would make the world best off were we to adopt it. But it is a live possibility that such a moral system might in some cases lead to individual actions that make the world worse off, at least in the short term. I acknowledge this problem: in the end, it may simply not be possible to truly have a theory that has all the advantages of consequentialism in their strongest possible form *and* which conforms to all our most reliable intuitions. Somewhere, something must give; this is merely where I have chosen to give.

But I still hold to be the core and motivating principle of my morality that what doing the right thing is *about* is making the world the best it can be. I endorse second order consequentialism not because I want to loosen this principle, but indeed because I think the best *way* to make the world as good as it can be is by applying consequentialism at the second order level, rather than endorsing this as a principle of *action*. It is for this reason, because this the core of how I think about morality, that I still call and think of myself as a consequentialist, even if

one might when viewing from a different angle very legitimately call my theory a form of deontology with a different guise on (as people sometimes say of rule consequentialism). It is also why, even though I think we have strong *independent* reasons to have a second order theory (which will be discussed in the next chapter but also occasionally show up in later chapters), I wish to see if I can make second order *consequentialism* work before I adopt some other second order theory.

Conclusion

In summary, consequentialism is to me a very attractive theory both for its emphasis and for its elegance. I am very drawn to the idea that what morality is ultimately about is making the world a better place. I also admire the way consequentialism has a single unifying underlying principle. It is because of the latter that I dislike modifications like satisficing consequentialism or agent relative theories that modify the goodness of consequences based on who is measuring it. These variants might avoid many of the biggest objections to consequentialism, but they do so at the price of sacrificing the relative simplicity that is one of consequentialism's main draws in the first place[16]. Instead, I believe that by applying the principles of consequentialism at the second order level we can avoid those same objections while retaining that elegance (at least at that level). Over the course of this dissertation I will demonstrate this with some examples, with particular focus on the problem of Demandingness. But first I will argue for second order theories in general, which is the subject of the next chapter.

[16] It might be argued that neither of these changes makes consequentialism much more complicated. Satisficing consequentialism, for instance, merely draws the line in a different place, it doesn't *add* anything to the theory. However, as we'll discuss in Chapter 6, this simple version of satisficing consequentialism runs into too many problems. Satisficing consequentialism can be refined to avoid those problems, but to do so it must in fact become more complicated than maximizing consequentialism. A similar pattern holds for agent-relativism.

Chapter 4: Why Second Order?

Introduction

In the last chapter, I made the case for consequentialism, and argued that a systemic approach to solving its problems is better than a piecemeal one. I believe that SOC is such an approach, and over the course of this thesis I will argue that it preserves the virtues of consequentialism while avoiding its main problems. In this chapter, I will discuss what I mean by *second order* consequentialism in more detail. I will argue that it is important to make the distinctions between orders explicit and firm, in the process explaining why my approach is better than the superficially similar approach of indirect consequentialism. I will discuss internalization and binding, two important concepts that highlight what I mean when I talk about different orders. I will argue that thinking about our reasons for acting supports a strong distinction between the orders. Finally, I will argue that one of the great benefits of a second order theory is that it allow us to interrogate our intuitions, and in fact that we need a second order moral theory of *some* kind if our ethics is to be complete.

Factors and foundations

In his introductory book on normative ethics, Shelly Kagan (1998) makes a distinction between what he calls *factoral* and *foundational* theories. When we evaluate a moral decision, we take into account all the morally relevant factors and weigh them against each other. A factoral theory tells us which moral factors are important and how to weigh them against each other. A foundational theory tells us *why* those factors are the important ones and why that way of weighing them together is the correct one. Another way of putting it is that the factoral theory tells us *how* to make moral decisions, while the foundational theory is concerned with *why* we

ought to make decisions in that way. Kagan makes this distinction because he believes that these levels are to a certain extent separable: that one can discuss much about normative factors, for instance, without committing oneself to a particular foundational view. Of course, a philosopher's factoral theory is certainly going to be informed by his foundational theory. Indeed, that is the point of foundational theories. But when it comes to philosophical theories in the abstract, we can discuss these kinds of theories largely separately of each other. In particular it is striking that factoral theories can look very different to the foundational theories on which they are based. Kagan himself has a paper (Kagan S. , 2002) where he posits a Kantian foundationalism that results in a consequentialist factoral theory. And of course, as we will discuss there are many kinds of consequentialist foundational theories that result in nonconsequentialist factoral theories, most famously rule consequentialism.

This distinction between factoral and foundational theories informs a lot of my thought. The consequentialism I wish to commit myself to is not a consequentialist factoral theory, but rather a consequentialist foundational theory. This does not necessarily mean I am not a consequentialist at the factoral level as well, but it is an open question whether foundational consequentialism in fact leads to factoral consequentialism. Much of the rest of this dissertation will be devoted to discussing precisely that question. For now I wish to commit myself to foundational consequentialism first and foremost. It is for this reason I call myself a *second-order* consequentialist.

When we look back to the arguments in favor of consequentialism from the last chapter with this framework, we can see that some push more for a foundational theory and some more for a factoral one. For instance, the simple but strong intuition that making the world a better place cannot be wrong pushes for factoral consequentialism. So does Kagan's argument that all

of us care about consequences and so if consequentialism can be shown to be sufficient for a complete moral theory than nonconsequentialism is unmotivated. By contrast, some of the other virtues of consequentialism I talked about, such as its simplicity, elegance, and unity of explanation, seem to be arguments primarily for foundational consequentialism. These are virtues of moral theories, and of *theories* in general, but it is harder to argue that they are virtues of decision making procedures. We want our decision making procedures to be *practicable*, but a simple theory may be less practicable than a sophisticated one. A simple decision procedure like 'act according to the best consequences' is in fact harder to practice than even a complicated set of rules of thumb. As we will discuss in the next section, act consequentialists are well aware of this. But it is still interesting to note that something can be a virtue at the theoretical or foundational level and yet not a virtue, and perhaps even a vice, at the factoral one. This distinction, between what makes a good justification and what makes for a good decision making heuristic, is the core idea I want to explore in this chapter.

Indirect consequentialism

In this section, I want to discuss the fairly common consequentialist strategy of *indirection*, and explain why second order consequentialism is, though superficially similar, a distinct approach. Indirect consequentialism is when act consequentialists advise following rules of thumb in ordinary situations because that is likely to lead to overall good consequences, but stop short of fully embracing rule consequentialism. I consider act consequentialism, which says that the morally best action is the one with the best consequences, to be a form of *first order* consequentialism. This is because it applies consequentialism *as* the criteria we use when we are determining what actions are right. By contrast, I consider rule consequentialism, which says that we should adopt and abide by those rules that lead to the best consequences for adopting them, to

be a form of *second order* consequentialism. This is because it applies consequentialism *to* the criteria we use for determining what actions are right (in this case, a system of rules).

But when it comes to decision making, both types of consequentialists will say that it is better to act according to rules of thumb than to weigh the consequences of each and every decision[17]. This leads to some misunderstandings: many act consequentialists are mistaken for rule consequentialists because they advocate following certain moral rules in most cases[18]. But this sort of *partial* rule consequentialism or *indirect consequentialism* simply advocates certain decision-making heuristics as being practical guides to behavior that are likely to lead to the best consequences. They are not saying that the *criteria* for what makes an action right is that it follows these rules, as a true rule consequentialist would.

The argument for indirect consequentialism is as follows: in day to day moral decisions we simply do not have the time to stop and consider all the possible consequences of each and every action we take. Indeed, doing so would lead to *worse* consequences than acting according to rules of thumb. In many cases the decision cost of sitting down to weigh all consequences precisely is too high. Furthermore, we are often unaware of the full consequences of our actions as we perform them. It is generally better in the long run to act according to a heuristic that we know to have had good consequences in the past than to act according to our imperfect knowledge of the future. Finally, there are aspects of human psychology to consider: allowing certain courses of action to be allowed for or encouraged in situations where they lead to good

[17] To go back to the very beginning, see e.g. Bentham (Bentham J. , 1789, pp. Chap IV, Sec. VI) "It is not to be expected that this process [his hedonic calculus] should be strictly pursued previously to every moral judgment."; and Mill (1861, pp. Chap II, Par. 19): "it is a misapprehension of the utilitarian mode of thought to conceive it as implying that people should fix their minds upon so wide a generality as the world, or society at large."; Or see also R.M. Hare (1981, pp. 46-47), who gives a detailed argument for a similar point.

[18] See, e.g. Shaw (2000) taking about how G.E. Moore is misunderstood in this fashion

outcomes might lead to people overstretching this justification and taking those actions even in cases where they would not, leading to worse consequences in the long run. These justifications for indirect consequentialism are largely empirical and contingent – they in many cases would not apply to creatures that were omniscient and had perfect judgement, for instance – but they are quite plausible when we are considering limited and imperfect humans. For instance, it is fair to say that in the majority of everyday situations that most people encounter on a regular basis, moral rules such as "Don't lie", "Don't cheat", "Don't steal" and "Don't kill" are the ones that lead to the best consequences (morally speaking), especially in the long-run or when there is uncertainty.

But this does not mean that the indirect consequentialist thinks that you should always follow such rules. Instead they are, to borrow a phrase from Tim Mulgan (2001), advocating rule-following *strategically*, because it usually leads to good outcomes. By contrast, the rule consequentialist *defines* the right action as one which follows the rules, and uses consequentialism to select those rules. This difference only becomes truly evident in cases where it is clear that breaking these sorts of rules would lead to better consequences: in such a case a rule consequentialist would stick to the rule and an indirect act consequentialist would break it. These are thus the kinds of cases that feature prominently in the philosophical literature around consequentialism, as they are the most useful for testing consequentialist intuitions.

Let us take the commonly discussed Transplant case. A doctor can kill one of his patients, harvesting his organs to save five terminally ill people. Should the doctor do this? It is a very strong intuition we have that he should not, and this is often used as an argument against act consequentialism, which would on one conception seem to argue that we ought to do so. It is not, however, necessarily the case that an act consequentialist should recommend that the doctor cut

up the patient. One common strategy to resist this conclusion is arguing that the long-term consequences of letting doctors cut up living patients are worse than letting people die for lack of organs. For instance, people would probably stop going to hospitals if this sort of thing happened in them, with far worse consequences overall. All the many reasons we discussed above for why one should still follow rules of thumb such as 'don't kill' in the majority of cases still apply – for instance, people are creatures of habit, and in the long-term it is better to inculcate in medical professionals a strict code of ethics that would prevent them from agreeing to cases like Transplant even if that single case considered in isolation would lead to good consequences. And so on and so on, there are in fact very many reasons why even an act consequentialism would recommend against chopping up patients in any *realistic* scenario of Transplant.

But if the thought experiment is refined to remove those reasons – if it is specified, for instance, that there is no risk in the surgery, that the doctor will not be discovered, and that the doctor knows both of these things with a high degree of certainty, that the doctor will never be in a position as to develop bad habits again, etc. – *then* many consequentialists are willing to bite the bullet on this case (Unger, 1996). One defense that can be given, which I mentioned earlier in the chapter on intuitions, is that our intuitions are designed for realistic cases. Of course in any remotely realistic case the doctor is prohibited from cutting up one to save five, but the suitably modified thought experiment in which he is not so prohibited is so *unrealistic* that our intuitions are no longer a reliable guide (Sprigge, 1965). To be clear, this is far from the only solution that consequentialists have proposed to answer this particular objection. Rather, my point is that indirect consequentialists do *not* hold that we should continue to follow rules of thumb when the factors that normally lead to them recommending us to follow those rules are no longer present.

Rule consequentialists, meanwhile, would advocate following those rules even in such cases, and that is the major difference between them and indirect consequentialists.

The cases that are common in philosophical discussion, however, are often common precisely *because* they are rare edge cases useful for stress-testing theories, and we should not forget that. In the majority of actual instances of moral decision making, even the most ardent act consequentialist generally recommends following a moral code rather than evaluating each decision on its own individual consequences. This makes for an interesting objection: while the difference between indirect and explicitly second order (like rule) consequentialism is real, is it that important?

Theory and practice

There is an interesting point to be made here: while there are *theoretical* cases where the rule consequentialist would differ from the indirect consequentialist, if those cases are only very unrealistic ones then could it be that there is in fact no *practical* difference between the two? Further, if they differ only in cases where our intuitions are unreliable, that would also seem to undermine some of the motivation for rule consequentialism. To recap, I think one of the main arguments for rule consequentialism, and an argument I made earlier in support of second order consequentialism generally, is that it allows for a theory that has many of the benefits of consequentialism (simplicity, unity of explanation, etc.), but has less unintuitive outcomes. However, if that only makes a *practical* difference in rare and unrealistic scenarios, isn't that argument undermined?

I think that it might be, in that if your primary motivation for moving from act to rule (or first to second order) consequentialism is that you are worried about cases like Transplant, then a

sufficiently sophisticated indirect consequentialism can assuage some of your worries. Whereas if you are worried about allowing for cases like Transplant even in *principle*, then perhaps not even rule consequentialism escapes problems. After all, it is possible to imagine a world where cases like Transplant have the best results if we allow for cutting up the patient even *generally*, though it is dubious that our own world is such[19]. If you object to cutting up the patient *ever* being permissible even in distant possible worlds, then rule consequentialism is also not going to be sufficient. In other words, the benefits of rule consequentialism for those worried about cases like Transplant are there only if you object to such cases in some types of unrealistic scenarios, but not others. It feels more usual that you are either worried about *all* scenarios or you are not worried about unrealistic scenarios at all, which would motivate nonconsequentialist foundational theories or indirect consequentialism respectively.

But at the same time, we should not be surprised that the differences between indirect and second order consequentialism are more theoretical than practical, for the virtues of consequentialism that SOC is trying to retain – i.e. simplicity, unity of explanation and so on – are precisely its *theoretical* ones. It may well be that there is little practical difference between SOC and some suitably sophisticated indirect consequentialism. But if two moral theories are otherwise similar – both cohering with our intuitions about all realistic cases, both evolving from simple and uncontroversial claims like that the right thing to do is to make the world a better place – then we ought, I argue, to prefer the one that has greater theoretical virtues, even if those virtues make little practical difference.

[19] I will discuss this objection with regards to my own theory in detail in chapter 8 (p. 136).

Of course, arguments from elegance or theoretical virtue are always tricky, for different people have very different senses for simplicity. An indirect consequentialist can easily argue that the sharp line second order consequentialists draw between the factoral and foundational levels is itself an unnecessary complication. But drawing such a sharp line aids the applicability of both sides of the line, allowing one to apply a relatively straightforward consequentialism when evaluating what factoral theory we should adopt, while similarly allowing us to apply that factoral theory without having to simultaneously keep in mind its foundational theory.

Furthermore, if we have *independent* reasons for drawing such a sharp distinction, then the claim that it is an unnecessary complication loses weight. As we discussed earlier, Kagan draws his distinction between factors and foundations because he thinks that factors can be discussed independently of foundations, and because people have the same factoral view with very different foundational views and vice versa. This independence is something like the independence of normative ethics from metaethics. When we are trying to create an ethical system or talking about a particular philosopher's views it is usually the case that the normative views are informed by the metaethical views and the factoral theory by the foundational one. But it seems we can discuss particular views in the *abstract* independently in each of the levels. If this is true (and it seems to me to be true, although I do think one can press this point), then this is at least indicative that this distinction is in some sense a 'real' one, rather than an unnecessary complication introduced by second order theories. However I think that this independence claim is a relatively weak argument (as indeed it is in the case of metaethics), and a better argument for making the distinction a sharp one has to do with our reasons for acting.

The problem of collapse: internalization and binding

It is often alleged that rule consequentialism 'collapses' into act consequentialism. There are a few ways this argument can be made, but the standard way the argument runs is as follows: rule consequentialism says that we should follow the set of rules that lead to the best consequences when followed. But surely that set is just the single rule "do the act that has the best consequences". For any other rule will *at best* have equally good consequences, and for most plausible rules there are going to be scenarios where it will have worse consequences. Thus rule consequentialism is really just act consequentialism with extra steps.

It is not entirely clear, though, that this constitutes a 'collapse'. I think that if we keep in mind that we are asking slightly different questions when we ask "what is the right thing to do" and "*why* is that the right thing to do", we see that this is not really a 'collapse'. Rather, it is a claim than any moral theory that is consequentialist on the second order level *must* also be consequentialist on the first order level – that the former inevitably leads to the latter. I think this is a slightly different thing than saying that rule consequentialism simply collapses into act consequentialism, though it is still a claim rule consequentialists want to deny.

How do rule consequentialists deny this claim? One of the most prominent defenses of rule consequentialism is laid out by Brad Hooker in his book *Ideal Code, Real World* (2002). Hooker argues that this objection only works if one judges the expected value of *complying* with rules, but that rather what we should be judging is the expected value of *internalizing* rules. To internalize a rule means to make it a part of your dispositions and judgements, to commit to following the rule in all cases, and to adopt it as a guide to your decision making. Hooker says that internalizing a rule has consequences over and above merely complying with it. A rule like "do only the action that has the best consequences" may be simple, but there are reasons why

internalizing such a rule might have bad consequences. For example, to go back to Transplant, if doctors only internalized that rule people would be too suspicious of their motives to go to the hospital etc. It addition, though it is simple in one sense it is quite a difficult rule to apply in practice, which means it has high internalization costs. So, for instance, it is more difficult to make part of one's disposition than a rule like, for example, 'don't kill'. It is also more difficult to teach and disseminate. All this calls into question whether SOC really would recommend act consequentialism on the first order level.

There's another way to make a second order consequentialist argument against act consequentialism at the first order level that Hooker gives a nod to, but which I wish to go into a bit more, and that is the idea of committing or binding oneself to a rule or principle. I think the easiest way to explain this idea is with promises. One can easily make a second order consequentialist argument for adopting the practice of promising, and for keeping your promises. Such an argument would include the benefits of not just developing a reputation for honesty but also the disposition to keep your promises, what Hooker calls internalization. But for the practice of promising to *have* the desired good consequences, it must be the case that one will keep the promise even if doing so would lead to bad consequences – within certain limits of course. Just as a legal contract cannot be enforced if it requires you to commit an illegal act, a promise cannot bind you to perform an excessively immoral act. However, it is plausible that for the practice of promise-making to be meaningful, you cannot abandon a promise if it will lead to *minor* bad consequences. Hooker calls this the difference between building in a 'prevent disaster' exception into the rules, which every plausible rule consequentialism must, and straightforward act consequentialism which would have one abandoning the rule *whenever* doing so would have better consequences.

It seems quite plausible to me that for the practice of promises to yield the best consequences it is necessary that we hold ourselves and others to them even if sticking to one doesn't lead to the best consequences – modulo exceptions for disaster of course. But from this it follows that we have a case where SOC does *not* yield act consequentialism on the first order level. Very similar things, I think, can be said of certain special obligations or relationships. If you think that the world is a better place when people have meaningful relationships with each other – whether because that makes people happy or because it is vital for human flourishing or for whatever other reason you might think so – then a second order consequentialist theory would advocate for such relationships. If you think that part of having such relationships involves taking on special obligations and responsibilities towards the people you have a relationship with, even if those obligations don't always tell you to take the action with the best consequences, then once again we have second order consequentialism not leading to first order consequentialism.

Reasons for action

There's another aspect of the distinction between orders that can be brought out with the examples of promises and special obligations, and that has to do with our reasons for acting. If you asked me why I keep my promises, I may well respond with a second order consequentialist justification, arguing that the practice of keeping promises no matter what is one that has very good consequences if adopted. Alternatively, I may give you a contractualist story, or perhaps a Kantian justification based on the categorical imperative, or so on. Presumably, I *have* some sort of reason for being a person who keeps their promises. But let us say I made a promise to aid a friend if they were in need, and the time has come to fulfill that promise. If I were to move to help that person, and you were to ask me why I was doing so, there are two ways I could answer.

And one of those ways – answering by giving you my consequentialist or contractualist or Kantian story – seems somehow *odd*. It's not that it's a wrong answer, indeed it seems correct as far as it goes, but there's a certain sense in which it doesn't seem to be the right answer either. The right answer would be "because I made a promise". That justificatory story I have is not actually part of the reasoning process that moves me to fulfill my promise. It is the promise itself that provides the necessary reason.

This argument is a nod to a famous argument by Bernard Williams (1981), regarding someone having "one thought too many" when taking action. Williams' example involves a man rescuing his wife, and how the wife would have reason to be upset if the man first thought about whether doing so was justified before he tried to save her. To quote: "it might have been hoped by some (for instance, by his wife) that his motivating thought, fully spelled out, would be the thought that it was his wife, not that it was his wife and that in situations of this kind it is permissible to save one's wife". A common response to this example is that Williams brought out a distinction between the reasons for acting in the moment of action and our broader moral theory. It has been argued (see, e.g. Lang & Heuer, 2012) that Williams himself would have been dissatisfied with this interpretation and in fact meant it as a critique of impartiality itself, but those same authors also admit that most people are unconvinced that it actually succeeds at motivating that more radical conclusion, as indeed am I.

Rather, what seems to be compelling about Williams' story is that there is a distinction between the sorts of reasons that guide our decision making and our justificatory story about those reasons. The latter implicitly underlies our decision making framework, and it is very important that it is there, but it isn't something that is *part* of our decision making process. This

might be called the distinction between proximate reasons and ultimate reasons, or between the factoral and the foundational, or, as I refer to it, between first order and second order.

And not only is there such a distinction, but it seems somehow *inappropriate* to bring the latter into contexts where the former is at play. It wouldn't just be *strange* were the husband to have one thought too many, but almost somehow wrong, as if to do so would cheapen the 'reason-givingness', perhaps, of his wife. This is connected into binding in a way, I think, or perhaps it is another sort of binding. When we enter into a meaningful relationship with another, particularly a spousal relationship, we have already commited ourselves to seeing that relationship as itself sufficent reason to aid them. The reason it seems inapproapriate to have that extra thought, I think, is because we are suspicious that a person that stops to think about the fact that they made a commitment every time it comes into play hasn't really commited themselves deeply, hasn't really, in Hooker's words, *internalized* the special obligation[20].

What I aimed to highlight this section and the previous one is that we have multiple reasons to make a distinction between the first and second order, and ones that are compelling independent of consequentialist concerns. This is important because, *pace* Hooker, I don't actually think that internalization and binding or similar notions can wholly defend rule consequentialism from the collapse objection. You may have noticed something about the arguments I have been using in favor of rule consequentialism in this section, and that is their similarity to the arguments I presented in favor of indirect consequentialism earlier. Someone like G.E. Moore (1903) will use some of the same arguments we used earlier and emphasize our poor epistemic position, arguing that we must follow proven rules even if it *seems* like violating

[20] I will come back to this argument in much more detail in Chapter 12 (p. 214), once some necessary groundwork has been laid.

them would lead to better consequences. Bentham was worried about the long-term consequences of allowing one to be sacrificed for many without restriction, which is why he built very strong protections for individual liberty into his theory, (1840) even though a naive interpretation of the 'greatest good for the greatest number' would have any sacrifice be justified if it benefited the greater good. Many act consequentialists have worried about exactly these sorts of things, and modified their theories to fit. I believe a sufficiently sophisticated indirect consequentialism can include internalization and binding and all the rest.

But I'm not very worried about this fact, because I *also* disagree with Hooker that if rule consequentialism and act consequentialism have the same practical implications then that is bad news for rule consequentialism. That only follows if rule consequentialism is introducing an unnecessary complication for no practical difference. But if the distinction between first and second order moral theories is one we want to be making *anyway*, then rule consequentialism no longer seems to be at a disadvantage. Indeed, if anything the tables seem to have turned, as it is that incredibly sophisticated indirect consequentialism that now seems overly complicated. That very sophistication has caused it to lose some of the straightforwardness that might have initially attracted one to act consequentialism. By constrast, by making the distinction between first and second order explicit and part of the theory from the get go, rule and other second order consequentialisms preserve some of that relative simplicity – even if only at the second order level. More importantly, indirect consequentialism makes this distinction between orders fuzzy and I think we have good reasons to want to be a sharp one[21]. For all of these reasons, I think that even though I'm pretty sure I could create an indirect consequentialism that is practically

[21] I will talk about how the lack of a sharp distinction here harms R.M. Hare's two level utilitarianism in Chapter 11 (p. 202)

equivalent, if I want to preserve the simplicity, unity of explanation and the other theoretical virtues that initially drew me to consequentialism then I am better served by developing a second order consequentialism rather than an overcomplicated act consequentialism.

Second-order theories and our intuitions

I believe that the moral system that SOC recommends will be an intuitively plausible moral system that resembles in most respects what is usually called 'common sense' morality. There are two reasons for this, the first of which is quite trivial but the second of which is much deeper. The trivial reason is that any plausible moral system, be it consequentialist or deontological, rule-based or act-based, will resemble common sense morality to some degree because that is the *measure* of plausibility. I find SOC to be promising *because* I believe it recommends a common sense first order morality. Some of the specifics of why I believe this will be apparent in later chapters.

For now I want to focus on the second, deeper reason that I believe that the recommendations of SOC coincide with our intuitions, and that has to do with the origins of our intuitions themselves. As we discussed in chapter 2, our intuitions do not arise from nothing but are rather shaped by many things: by evolution, yes, but also by upbringing, by culture, by experience, by knowledge of history. I do not bring this up, as other consequentialists do, to cast doubt on them. Indeed, I believe the opposite is true, that this gives us a reason to be less skeptical of them. Our moral intuitions reflect in part a sort of accumulated wisdom, both our own and that of the societies in which we live. This wisdom is inexact, and of course there are good reasons to believe that our intuitions are likely to be flawed in some important ways, but at the same time it deserves to be taken seriously. And part of this wisdom is the knowledge of

what the past consequences of certain actions are, the knowledge of which moral principles have been productive for human good and flourishing and which have not.

I hasten to add that it should not be taken from this that I intend to argue that all our intuitions are implicit second order consequentialist reasonings well disguised. I judge that this is far from likely. But there is surely an element of second order consequentialist reasoning that underlies some of our moral intuitions. Even my so called 'pre-theoretic' moral judgements are informed by my knowledge of history and experience of the world, including my knowledge of the consequences of taking certain actions and adopting certain principles. This even holds true for those moral intuitions that are the result of evolution, as evolution is in a sense 'consequentialist', in that we would only have evolved some particular moral intuition because of its good adaptive consequences. The kinds of consequences evolution cares about are very far from moral consequences, but it is not like there is no overlap. When we take all these things into account, it seems to me that while we should not expect there to be a perfect concordance between the results of a considered second order consequentialist theory and our intuitive moral principles – and indeed, we should expect there to *not* be such a perfect concordance – we should expect there to be a resemblance.

This deeper reason also highlights another way of understanding the purpose of second order moral theories. Just as a first order moral theory is an attempt to create a more coherent, complete and well-grounded refinement of common sense morality, so too is a second order theory an attempt to do the same with our moral intuitions. This means that a second order theory provides us with a way to examine, critique, refine and potentially revise our intuitions, which is particularly important in cases where our intuitions conflict with each other or where we have good reasons to believe that they are particularly untrustworthy or flawed in some critical way.

This ability to critique and potentially correct our intuitions is one of the main reason why I think we need a second order moral theory, apart from the reasons already discussed earlier.

Second order consequentialism and our intuitions

Second order consequentialism is a theory that refines our intuitions, but that very fact opens it up to criticism. We ought to be cautious of any theory that asks us to abandon or replace deeply held moral principles, but the mere fact that SOC does so is not in itself a real criticism. This is because any other plausible second order theory would ask us to do the same. Pure intuitionism, the idea that we ought to take our intuitions as given without critique or reflection, is in a sense a second order moral theory, just as the idea what we simply act according to what feels right in every scenario is in a sense a first order moral theory. But they are both rather poor and implausible theories, because our intuitions sometimes conflict and are not always clear, and because we have good reason to believe that they are flawed in some systemic ways. So the mere fact that SOC is revisionary is not a real problem.

The more serious issue is that we have some reason to think that SOC is likely to be far *more* revisionary compared to other second order theories. This is for the same reason that first order consequentialism is generally seen to be more nonintuitive than other first order theories. Consequentialism must reconstruct everything in consequentialist terms, but many of the moral beliefs that we hold dear seem very nonconsequentialist. It is not that such reconstruction cannot be done, as in most cases it can, but that doing so is often unsatisfactory to some. For instance, first order consequentialism reconstructs the rule ‘do not murder’ as ‘because murder almost always results in bad consequences, and because we often operate under uncertainty and incomplete information, it is better to not commit murder in all but the most extraordinary circumstances’. The practical implications for action are generally quite similar, but one might

reasonably be unsatisfied that this actually captures everything we want to capture about the moral belief in question. Similarly, even if SOC results in a plausible and attractive first order moral theory, one might reasonably be skeptical of whether it does so for the right reasons, and whether consequentialism even at the second order level captures what we want out of our moral principles.

To this charge I can only say that I cannot defend second order consequentialism in the abstract. I believe that the revisions SOC demands of our moral principles are deserved and that the costs of modifying them in such a way are outweighed by the benefits of a consequentialist theory at a second order level. But I cannot make that argument without first showing what those costs and those benefits are. I hope I have managed to convince you that some form of second order theory is necessary, and even that most moral theorists already have one even if they don't make the distinction between orders as clear as I or Kagan do. If you are attracted, as I am, to consequentialism in any form, whether because of its simplicity or unity of explanation or because you are drawn to the notion that the best response to values is to maximize them, then it is worth investigating the possibility of consequentialism on the second order level just as it is at the first. My contention is that SOC is much less nonintuitive, less revisionary, than consequentialism is at the factoral level. But the only way to prove this is through demonstration.

Conclusion

I argued in this chapter that the distinction I draw between first and second orders is a necessary one for moral theorizing. If we do not make this distinction an explicit one, we fail to capture how our reasons for action may be quite different from the ultimate justifications for those reasons – the latter of which have no part to play when we actually *act*. I also argued that a second order theory gives us a way to interrogate our intuitions and a framework for balancing

conflicting first order principles. Many times over this dissertation we will encounter cases where without such a framework we are left with a deep problem of arbitrariness in our moral theories. The case that consequentialism, specifically, makes for a good second order theory cannot be made in this chapter alone, and is the business of the rest of this dissertation.

Chapter 5: What is Second Order Consequentialism?

Introduction

In the last two chapters we went over reasons why one might be attracted to consequentialism, as well as why we should draw a sharp distinction between first order ethics and second order ethics. Putting those together gives us reasons to explore the possibility of second order consequentialism, and the rest of this dissertation will be an exploration of what I believe to be the most plausible version of SOC. In this chapter, I will clarify what I mean by second order consequentialism specifically, and what I mean by adopting and internalizing a first order theory. I will also discuss the similarities and differences between my theory and rule consequentialism, and why I think that many objections to the latter do not apply to my theory.

Why 'second order' rather than 'rule'?

Throughout this dissertation so far I have largely discussed rule consequentialism, using it as a stand-in for all forms of second order consequentialism. So why do I not simply call myself a rule consequentialist, rather than insisting on this idiosyncratic term I have invented? The simple answer is that I am not wedded to the idea that the correct sort of first order theory is a system of *rules*. That doesn't mean I *don't* think that sometimes the correct first order principles are the kind that standard rule consequentialism recommends (i.e. a set of rules). Rather, the claim I am interested in and want to defend is the more general claim that the correct sort of second order theory is a consequentialist one. It may be that it is not really possible to have a single even relatively simple first order theory, for as we saw in the last chapter even act consequentialism must develop into a complicated system of indirect consequentialism if it is to be a plausible system of ethics. The world might simply be too complicated for us to have

maximizing consequentialism on the first order level, but I think there is still hope for it on the second. That is the idea that the rest of this thesis will be dedicated to exploring.

Let me start my laying out in brief what my theory looks like in broad strokes. Second order consequentialism is in essence the claim that we ought to judge the moral principles we live our lives by on a consequentialist basis, and live by the one(s) that will have the best consequences in the long run if adopted. The purpose of SOC is not to provide criteria for moral decision making or to be action guiding, but to analyze and interrogate the underlying moral principles we use for that end. It is deliberately silent about the content or even the *form* of those moral principles (what I call first-order theories), which is what distinguishes it from rule consequentialism. By first order theory I do not necessarily mean a complex or abstract moral system as might be implied by the term 'theory' but simply whatever means by which one decides which action is the most moral in a given circumstance. I simply use 'theory' because there's no other good catch-all term, but even just 'do whatever comes to mind first' is a first order theory by my terms (though not a good one or one I think SOC would ever recommend).

It is plausible that in most cases SOC *would* endorse a rule based first order moral theory, for the same reason that indirect consequentialism recommends following a set of rules. And indeed, this is also the same reason that rule consequentialism and other rule-based moral theories are attractive: because a moral system based on rules is a good practical guide to decision making that is also intuitively attractive to us. However, a rules-based system is not necessarily the only kind of decision-making heuristic that is practical and intuitively attractive, and nor is it always going to be the one that leads to the best consequences in the long run if adopted, so my theory does not fully converge with the rule consequentialism of Hooker and others. In particular, as we'll discuss at length in chapter 13, rule based systems of morality fare

poorly when used to govern how we deal with close personal relationships, and this is one of the things that motivates me toward a pluralism of first order theories. The last few chapters of this dissertation (11-13) are dedicated to arguing for such a pluralism.

Another reason I would rather not think of first order theories as only being sets of rules is that it obfuscates one of the main considerations that second order consequentialists have to keep in mind – that of internalization. *Humans are not robots*; we cannot simply insert a new line of code – "do X in circumstances X" – into our programming. Internalizing a moral principle may sometimes involve agreeing to be bound by some rule, but if that *alone* were enough to make us comply with that principle then no one would be out of shape or procrastinate at work. Often, to really internalize a first order theory we also have to inculcate certain habits in ourselves, or develop dispositions that cause us to react in certain ways. As we'll discuss below, some of those dispositions may be to have certain emotional reactions rather than to act in certain ways. You *can* think of all of these as rules that we are following (Hooker does) but it seems to me that in many cases – such as with the example of close personal relationships – doing so would be somewhat misleading.

At this point, the rule consequentialist may argue that any first order theory, or combination thereof, can be formulated as a set of rules, and therefore that any kind of second order consequentialism can be easily recast in rule-consequentialist form. This seems to me to be true, but only in a trivial sense. To give an analogical example, it is technically correct to say that classical utilitarians are rule followers of the single rule 'do that action which maximizes utility' but it would be extremely unhelpful to therefore argue that classical utilitarianism is a form of rule consequentialism with one rule. To do so would simply muddy the waters. In general, and we will also discuss this when it comes to virtues (Chapter 13), every moral theory performs

some of this sort of recasting to some extent. Rule based moral theories reconstruct virtues as a kind of rules, indirect consequentialists reconstruct deontological constraints like ‘don’t murder’ as rules of thumb, etc. The difference is what they give *primacy* to. I do not give primacy to rules at the first order level: when it comes to evaluating first order theories SOC only analyses them as being a set of rules if doing so makes more sense than otherwise, and I’m not convinced that that is generally true.

For this reason, I prefer the more neutral term of ‘second order consequentialism’, and consider rule consequentialism to be a *subtype* of SOC that *primarily* views its candidate first order theories as sets of rules. I view this as similar to the distinction between consequentialism and utilitarianism. The utilitarian argues that all the goods we care about can be instrumentally derived from wellbeing, but most consequentialists don’t argue for that, either because they don’t agree or simply for reasons of time, expedience, or interest. Consequentialism can be and often *is* defended independently of arguing for or against any particular theory of the good (Kagan S. , 1998). Similarly, while traditionally SOC has been discussed almost entirely in the context of rule consequentialism, it is my contention that it can be defended independently of arguing for or against particular organizing principles (i.e., rules vs virtues etc.) of first order theories – and that kind of defense of it is what I aim to provide here. To argue for rule primacy at the first order level would require *additional* arguments that I am not here interested in making. In making the argument for the more general SOC instead of specifically rule consequentialism, it is my aim to make fewer claims, not more.

Global Consequentialism

At this point is it also worthwhile discussing another approach to consequentialism that embraces neither acts nor rules: *global* consequentialism (Pettit & Smith, 2000). According to

Pettit and Smith, unlike a *local* consequentialism such as act or rule consequentialism, global consequentialism does not privilege any particular category of evaluand. Instead, it simply asks that we seek the best outcomes in every measure. Pettit and Smith claim that *every* form of local consequentialism allows for situations that clash with our basic *consequentialist* intuitions, and therefore will fail on consequentialist terms rather than being merely generally nonintuitive. The basic idea is to imagine a world with a super powerful mad scientist or evil god. This entity punishes traditionally positive examples of a given evaluand but cares not for others. So for instance: in the case of local motive consequentialism the mad scientist will ensure horrific consequences unless people hold malevolent motives, but he cares not for the actual acts that people perform. Local motive consequentialism would then have to conclude that the right motives to have in such a world are the malevolent ones. But then it also seems committed to saying the right acts are those motivated by malevolence. However, it seems more in accord with our pre-theoretic intuitions to say that the presence of the mad scientist would mean that the right thing to do is to have malevolent motives (as otherwise he will punish millions of people) but to then *not act on them* (because the mad scientist does not care how we act), acting instead to maximize utility.

A similar argument can be made for any local consequentialism (e.g. act, rule, or virtue consequentialism) and, therefore, if one wants to be a consequentialist, according to Pettit and Smith one ought to be a global one. Whereas if scenarios like the mad scientist one lead you to doubt that the right thing to do is maximizing value then, well, you ought to not be a consequentialist at all – this is intended to be a consequentialist argument for global consequentialism. I am actually less impressed with this particular line of argument – I am not hugely convinced of the argumentative value of such fantastical scenarios. What I am drawn to is

the general claim that if you really care about maximizing value above all else, then in scenarios where a judging consequences by a single evaluand results in a bad consequences by every other evaluand you should abandon your commitment to that evaluand. That is part of the reason that I am silent about whether my first order theory evaluates on the basis of rules or motives or so on. I want to evaluate moral theories holistically, and ask which one will have the best outcome *overall* when we adopt it. In other words, I do not want to privilege any particular evaluand *in principle*, though as we'll discuss below and in the rest of the dissertation SOC may in practice recommend first order theories that privilege one evaluand over others. Therefore, I do believe in global consequentialism, as Pettit and Smith describe it, at the second order level (but not necessarily at the first order level).

This might lead to a concern that I can no longer quite as easily claim that my theory can lead to nonconsequentialism at the first order level. A rule consequentialist can point to internalization and binding as a reason to favor rules rather than acts at the first order level. But a second order consequentialist who favors *any* evaluand at the first order level seems vulnerable to a Pettit-Smith style objection where it is easy to construct a scenario where that evaluand leads to outcomes that are bad as measured in the terms of all the other evaluands that the second order consequentialist also cares about. It might seem then that second order global consequentialism leads to global consequentialism at the first order level as well. At that point, one might wonder why it is necessary to make such a distinction in the first place, and whether introducing the ideas of orders is necessary when one can simply be a global consequentialist, evaluating everything in terms of the best overall consequences.

To this I have two things to say. The first is that, as I said in the last chapter, I think we have *independent* reasons for making the first/second order distinction, and that there are

theoretical virtues to doing so even if the practical differences between a second order global consequentialism and a first order global consequentialism are trivial. The more interesting response is the second one: I think it is not so obvious that global consequentialism at the second order level is vulnerable to this sort of objection, so long as we are willing to consider that a first order theory can have limited scope. If so, then the Pettit-Smith scenarios need not be troubling, so long as such scenarios cannot occur within the specified scope of each first order theory. This opens up the possibility that a single second order theory might recommend multiple different first order theories for different circumstances, what I call first order pluralism.

Of course, we need more reason than simply avoiding SOC collapsing into first order global consequentialism to believe in first-order pluralism. The last few chapters of this dissertation will go over in detail why I think we have good reason to be pluralists. In chapter 11 I will discuss how many deontologists already believe in thresholds above which deontological constraints may not apply – e.g. that it may be permissible to kill to save a million people – and thus implicitly accept a limited scope for those constraints. In Chapter 12 I will argue that we have good reasons to use different sets of moral principles for small-scale cases than for large-scale cases. Finally in chapter 13 I will argue that the sorts of moral principles we use to deal with strangers are not appropriate when used for close personal relations, and again argue for different sets of moral principles for different cases.

If we do accept the idea of first order pluralism, then being a global consequentialist at the second order level – which I am – does not necessitate embracing global consequentialism at the first order level as well. Instead, SOC will recommend different first order theories depending on the circumstances. The presence of a Pettit/Smith style mad scientist will of course influence *which* first order theories SOC recommends, but so will any number of other

contingent factors. This is another reason why I do not consider myself a rule consequentialist: even when SOC *does* recommend that we adopt a system of rules at the first order level, it does so because of contingent facts about the world. In a different world, it might recommend differently.

What it means to adopt and internalize a moral theory

Earlier I characterized SOC as the claim that 'we ought to judge the moral systems we live our lives by on a consequentialist basis, and live by the one(s) that will have the best consequences in the long run if adopted'. In this section I want to elaborate on what I mean by 'adopting' a first order moral theory, and expand a bit on the ideas of internalization and binding.

Adopting a first order moral theory means binding or committing to using that theory as a decision-making procedure for those areas where it applies (as we'll discuss in the last few chapters, a first order theory may have limited scope). In particular, it means that we should follow the first order theory even in cases where the principles underlying our second order theory might recommend differently *in that individual case*. In the case of SOC, that means we commit to following the first order theory even if doing so would lead to less than the best consequences in some limited circumstances. That said, it is worth noting that the first order theories that SOC recommends will plausibly have 'prevent disaster' exemptions that will advise us not to follow them if doing so would lead to unacceptably bad consequences. It is also worth noting, as we'll discuss below, that SOC may in some cases recommend act consequentialism at the first order level as well. Even in this case, we still bind ourselves to the first order theory: while in this case we would do that act which would lead to the best consequences, we would do so because we have bound ourselves to a consequentialist first order theory and not because our second order theory happens to be consequentialist. Both theories happen to be consequentialist

in these cases, but they are still operating at different levels. Nonetheless, in those cases where SOC *does* recommend a nonconsequentialist first order theory, we would follow what that theory says when making decisions even if doing so would not lead to the best consequences in those limited circumstances.

Internalizing a first order theory means incorporating it into your dispositions – not just using it as a guide for your decision making but training yourself to actually abide by it when you need to make a decision. Internalization is necessary because, once again, humans are not robots: binding ourselves to a set of rules is more complex a process than simply memorizing the rules and making a mental note to abide by them. Or at least, it must be if we want to actually *abide* by those rules when push comes to shove. Further, acceptance of a first order theory doesn't just mean accepting to follow the principles of the first order theory for yourself, it also means, to quote Hooker, "the disposition to encourage others to comply with them, dispositions to form favorable attitudes toward others who comply with them, dispositions to feel guilt or shame when one breaks them and to condemn and resent others' breaking them, all of which dispositions and attitudes being supported by a belief that they are justified." (Hooker, 2002, pp. 76-77). All of this means that *adopting* a first order theory has consequences over and above the consequences of simply complying with it.

To provide a concrete example of these general ideas, let us return to the perennial case of Transplant. Why do I believe that in these circumstances the doctor should not kill the one patient so that they might harvest the organs to save five more? It is because I believe that the set of first order principles that would have the best long term consequences if the doctor were to adopt them would prohibit such actions. Indeed, as part of training to become a medical professional, doctors condition themselves to simply rule out certain courses of action from

being legitimate options, and instill in themselves the disposition to react with horror and disgust[22] to the idea of deliberately killing an otherwise healthy patient or harvesting their organs without consent. That is not to say there is no room to question these principles or change them – one might argue, for instance, that mainstream medical ethics (in the West) privileges patient autonomy more than it ought to. But the time and place to make these discussions and adjustments is not *in the operating room*. It is, rather, the slow process of trying to evolve a more accurate medical ethics over time that takes place when the doctor is not having to make critical, moment-to-moment decisions. One of the primary reasons to make a distinction between orders, as I discussed in the last chapter, is to separate these kinds of discussions out from the kinds of considerations – first order considerations – that guide our decision making. When actually operating on a patient, the doctor should follow the rules they are committed to and not debate those rules. As for the reasons why they shouldn't, we went over them in the last chapter (decision costs, uncertainty, etc.)

As I've framed it, one of the main reasons we internalize first order theories is that it makes it more likely we will actually comply with them; this naturally raises the question of what sorts of compliance rates I envision for my first order theories. In fact, I do not have any particular level of compliance in mind, but rather feel this is one of the things that the second order theory needs to weigh in its calculus. For example, it seems to me that SOC would not recommend a first order theory that would result in the best world if everyone was to comply with it if said theory was such that we expect no one to actually comply with it. This is because even though the first order theory theoretically has good consequences its *actual* consequences

[22] Incidentally, that SOC might recommend this sort of thing – instilling in ourselves a disposition to react with horror to certain courses of action – is one reason why I prefer not to call myself a rule consequentialist. It's not that you *can't* think of that as a rule ("react with horror if – ") but that I find it unhelpful to think of it like that.

are poor. But that should not be taken to mean that the SOC will recommend only those theories that are easy to internalize and comply with. It seems likely that a theory will only have near perfect compliance if it is tremendously undemanding and asks that we sacrifice almost nothing. A first order theory that asks more of us (and which therefore will have poorer compliance rates) might nonetheless have better overall consequences if adopted, in which case SOC will recommend this latter theory over the one with 100% compliance.

In other words, compliance as well as rate of adoption – i.e., how easy it is for someone who has internalized the theory to get other people to internalize the theory as well – are two factors among many that go into evaluating the consequences of adopting a first order theory. Crudely, we might multiply the expected consequences of the theory at 100% compliance with the expected rate of compliance to get the 'compliance-adjusted expected consequences', but in practice I expect the actual calculation to be rather more complicated than that. For example, a theory might have good consequences if widely adopted, but bad consequences if not (we'll discuss examples below) – which speaks against such a straightforward calculation. I do believe, however, that the first order theories SOC plausibly recommends will be of the sort that will be generally, but not universally, complied with.

It seems to me that Hooker is reluctant to take this approach because he wants the first order rules to be both the guide to our actions *and* the criterion by which we make judgements. To quote him:

> It seems to me counterintuitive that what is morally right depends on rules designed on the assumption that we will regularly fail to comply with them. If the point of setting a rule one place rather than another is that our actions will miss their target to some degree,

> then a human tendency to make mistakes is shifting the line between the morally allowed and the morally forbidden. (Hooker, 2002, p. 77)

I, on the other hand, already to some degree separate how we determine what is morally right to do and how we decide when it is appropriate to blame others (as we will discuss more in Chapter 8). In addition, I am not wedded to using the set of rules generated by SOC as the criterion of right action (this is perhaps another way in which I am not a rule consequentialist). Finally, I am less troubled by the idea that our moral judgements may be based on standards we do not expect everyone to comply with, and to some extent I even think it desirable. Part of the purpose of moral judgements is to exhort us to be better than we are – again, we'll come back to this point in chapter 8.

Individual and Collective Consequentialism

There remain further clarifications to be made to SOC to explicate it fully; in particular, I should clarify whether I am deploying it in its individual form or its collective form. There are two ways of interpreting SOC as I've laid it out so far.

1) I should adopt and internalize that first order theory (or theories) for which it is true that my adopting them leads to better consequences than if I adopted any alternative first order theory; you should adopt and internalize that first order theory (or theories) for which it is true that *you* adopting them leads to the best consequences; and similarly, anyone else should adopt that first order theory for which it is true that *their* adopting it leads to the best consequences, etc. (Individual Form)
2) Everyone should adopt and internalize the first order theory (or theories) for which it is true that *everyone* adopting and internalizing this set of moral principles would lead to

better consequences than everyone adopting and internalizing any alternative set. (Collective Form)

I will borrow terminology from Mulgan (Mulgan, 2001) and refer to these as Individual Consequentialism and Collective Consequentialism, respectively.

Thus far, I have been rather free with the usage of such language as "*we* ought to judge the moral systems we live our lives by on a consequentialist basis" and similarly free usage of the plural. It may therefore come as a surprise when I say that I in fact favor the Individual form and not the Collective form of my theory. However, I would hasten to add that there are many caveats to the most plausible form of Individual SOC that, in my view, license me to be so free with the usage of the plural.

The first of these caveats has to do with the *processes* by which we come to decide which first order theory has the best consequences. As we discussed in chapter 2, our reflective intuitions are not merely our gut feelings but are produced through a process of refinement that includes discussion with and challenges by other people. Similarly, our moral theorizing also involves taking input and arguing with others. Thus, even if I am trying to decide which first order theory is the one that leads to the best consequences if *I* were to adopt it, the means by which I come to that decision includes the arguments and insight of people other than me. Philosophy is by its nature a collective enterprise, and when I say things like "the first order theory that we estimate to have the best consequences" that *we* is more of an acknowledgement of the collective nature of this enterprise than an indication of whether the theory itself is in an individual or collective form. This is a relatively minor point, however.

The more important caveats have to do with the *kinds* of first order theories that I believe a plausible SOC would endorse. I argue that there are a number of constraints operating on such theories that makes it the case that even though I am adopting that set of principles for which it is true that *my* adopting them leads to the best consequences, and you are adopting that set of principles for which it is true that *you* adopting them leads to the best consequences and so on, nonetheless SOC would recommend that we adopt the same, or at least a *broadly similar* set of principles. This, in turn, is why I often talk about the principles that SOC recommends 'we' should adopt: this should be taken as shorthand for the overlap between set(s) of principles SOC recommends each individual should adopt – a shorthand I feel licensed to use because I believe that overlap to be considerable. There are a number of constraints that generate this substantial overlap.

The first of these constraints is *universalizability*: the moral principles recommended by SOC to one person must be the same as the moral principles SOC would recommend to another person in exactly similar circumstances. This should go without saying, and should simply follow from the fact that the consequences SOC is weighing are assessed from an impartial point of view (which, to be clear, is true of my version of SOC). But what this means is that it *can* be said that you and I (and Deepika and John) are following the same set of moral principles provided those principles are appropriately conditional – meaning that if I were in your position I should do what you should do and vice versa. While universalizability mostly falls out of the fact that SOC takes the impartial point of view, there are also independent arguments that the first order principles we follow ought to be universalizable. I am not a Kantian, but I nonetheless do feel some attraction to the argument that principles that can be willed to be universal law are at least preferable to those that cannot, all else being equal.

The remaining constraints all have to do with the types of moral principles human beings can effectively internalize. Like Hooker I think that in order to internalize a moral principle you need to *believe* in it, and therefore be motivated to act in accordance with it, as well as to judge others by it etc. Universalizability matters here too, since it's a lot harder for me to genuinely believe that a moral principle is right if I don't believe that another person in exactly similar circumstances would also be bound by the same principle. Similarly, internalizing a principle means endorsing it, being willing to defend it if pressed, judging other people by it, and encouraging other people to comply with it as well. That last point is particularly important, since it means that the *consequences* of me adopting a first order theory are not limited to me alone. If I successfully encourage other people to act according to the theory then the consequences of them adopting the theory are also included in the consequences that SOC is assessing when judging first order theories. That is why I said above that rate of adoption is also one factor (among many) that is taken into account by SOC.

Given all this and that even in its 'individual' form SOC takes the actions and reactions (or rather, the consequences of such) of other people into account, one might question whether I am actually arguing for the collective form of SOC after all. However, with the possible exception of universalizability, all of these factors are *contingent* ones. Essentially, insofar as my version of SOC tends towards the collective form it is because of features of human psychology more than anything else. Another way of looking at it is that I don't think humans are different enough (in *morally relevant* ways) for SOC to recommend first order theories that are significantly different across individuals. That doesn't mean the first order theories SOC recommends aren't relative to the individual at all: as I will note in Chapter 9 (pp. 143-144.) the first order theories that govern our judgements of blame certainly relativize by species at least.

And there *are* some cases where I do believe that the first order theories that SOC recommends are going to be very different from individual to individual, especially when it comes to the principles governing the way we interact with close personal relationships, which depend on details that are highly, well, personal. We'll discuss this case in detail in Chapter 13.

In summary, I prefer the individual form of SOC rather than the collective form, but most of the time I use expansive terms like 'the principles SOC says we should adopt' because I believe that in *most* cases the set of moral principles that SOC says I should adopt is substantially similar to the one SOC says you should adopt (or Deepika or John). But that is because of factors that do not apply to all cases (and as we'll see in Chapters 11-13, I think SOC recommends different sorts of first order theories in different cases) nor do they apply in certain imaginable scenarios where some of these contingent factors are either missing or overridden by more pressing factors. We've already mentioned a Pettit/Smith style 'mad scientist' scenario which could cause that, for instance, and we'll discuss some more scenarios below.

But there is one more major reason to prefer the individual form over the collective form, and that is that the collective form of SOC is vulnerable to a major and influential objection: The Ideal World Objection. An objection with needs addressing since it might be thought to apply to my theory as well, though I will argue it does not.

The Ideal World Objection

The Ideal World Objection (IWO) (Parfit, On What Matters: Volume 1, 2011) is generally taken to be one of the premier objections to Rule Consequentialism, but I dispute this characterization. IWO, I will argue, is really more an objection to Collective Consequentialism. Now, these are sometimes taken to be synonymous: Mulgan, for instance, argues that Rule

Consequentialism is a *form* of Collective Consequentialism (Mulgan, 2001, pp. 53-54). I disagree with this categorization, because I believe not only can you have forms of Collective Consequentialism that are not Rule Consequentialist (a point not in dispute, as Mulgan himself provides many examples (*Ibid.* pp. 104)) but that Rule Consequentialism also need not involve Collective Consequentialism. In my view, Rule Consequentialism is a form of Second Order Consequentialism for which the candidate first order theories are primarily in the form of rules, and SOC is simply consequentialism applied at the second order level. SOC can, as I noted in the previous section, come in both individual and collective forms.

There are, it seems to me, two main reasons why Rule Consequentialists typically (Hooker, 2002) (Brandt, 1992) (Parfit, On What Matters: Volume 1, 2011) advance the collective form of their theory: the Demandingness objection and the Collapse objection. The former will be discussed in great detail in later chapters (6-9), but suffice to say for now that I do *not* use collective rules as my means of answering the Demandingness objection and indeed agree with many of Mulgan's critiques (Mulgan, 2001, pp. 67-87) about using Collective Consequentialism to solve the objection (I'll come back to this in Chapter 7, pp. 117-122). With respect to the Collapse objection, we discussed in the previous chapter why I am not worried about it. Given this, there is little incentive for me to take the Collective form of SOC, so many or most forms of IWO do not apply to me.

The most basic form of IWO (Parfit, On What Matters: Volume 1, 2011, pp. 312-320) considers the issue of rules that would lead to the better consequences than any alternate ruleset if everyone complied with them but bad consequences at anything less than 100% compliance; Parfit gives the example of extreme pacifism. To deal with this form of the objection, rule consequentialists (including Parfit himself) relax the assumption of perfect compliance (Brandt,

1992) (Hooker, 2002) (Parfit, On What Matters: Volume 1, 2011), to various degrees. However, Abelard Podgorski (Podgorski, 2018) argues that this still leaves them vulnerable to the more general form of the IWO, which he calls the *distant worlds objection*.

Podgorski argues that the problem with rule consequentialism was never ideal compliance, but that the rules are formulated by referral to worlds that differ from our own in more than just our actions. What this means is that we can *always*, even for imperfect compliance, generate a world where following a set of rules which are optimal for any given level of compliance but which generate awful consequences for any other level of compliance. As for Parfit's suggestion that the correct rules are those that are optimal for *every* level of compliance, Podgorski argues that we can easily guarantee that no such set of rules exist by setting up the worlds so that the rules that have the best consequences at one level of acceptance are incompatible with those that have the best consequences at another level of acceptance (*Ibid.* p. 9).

Podgorski argues that in order to escape the distant worlds objection, rule consequentialism has to consider the consequences of the agent following the rules in their actual, maximally specific circumstances C. But to do so is to risk the Collapse objection:

> If, alternatively, the consequences of the agent's compliance in the actual world *do* matter for a rule's evaluation at some level of compliance, the view collapses into act consequentialism. To see this, let S be a set of rules that is best at every level of compliance, or best on average. Suppose S recommends that I do something other than what AC recommends in my maximally specific circumstance C. Then the rules S*, which say "Do as S recommends, except in C, do what AC recommends", would have better consequences than S at the relevant level of compliance, because complying with S and complying with S* overlap for every case except in C, where S* does better. At all other worlds and levels of compliance, S and S* are evaluated identically. Since S* does better than S at one level of compliance, and the same everywhere else, S does not have the best consequences at every level of compliance, and does not have the best average consequences across all levels. By contradiction, S cannot recommend that I do something inconsistent with AC. (*Ibid.* p. 10)

But, again, I am not worried about the Collapse objection. Indeed, my version of SOC is very much concerned with the consequences of the agent following the first order principles in their actual, maximally specific circumstances. The result of this is that my theory *will* recommend following act consequentialism if doing so is what would lead to the best consequences overall. Chapter 12 is dedicated to exploring the exact possibility that we should be more act consequentialist in some situations and more nonconsequentialist in others. In fact, "Do as S recommends, except in C, do what AC recommends" is exactly the sort of first order pluralism that chapters 12 and 13 explore and argue for.

Implicit in this argument – that my version of SOC recommends that we sometimes be act consequentialist – is that I must concede that one can set up possible worlds where SOC would recommend that we are act consequentialists in *all* situations, and I do indeed concede this. Even in such a world, though, I would *still be* a second order consequentialist – I would merely *also* be always an act consequentialist at the first order level (as opposed to what I actually am, which is only sometimes an act consequentialist at the first order level). To the extent that SOC recommends nonconsequentialist first order principles (and I do believe it often, but not always, does so) it does so because of the *actual* consequences of adopting those principles in our own maximally specific circumstances, not by referencing the consequences of adopting those principles in another world at some level of compliance. In this way, my version of SOC is not vulnerable to the distant worlds objection, either in its general form as presented by Podgorski or the specific case of the Ideal World Objection.

Conclusion

So is SOC more similar to rule consequentialism or to act consequentialism? That depends a lot on your view of rule consequentialism. If, like me, you think the most important

thing about rule consequentialism is that it is applying consequentialism at the second order level to *select* the principles we adopt and internalize rather than *as* the principles we adopt and internalize than you will see rule consequentialism as a form of of SOC. If, on the other hand, you think Collective Consequentialism is a core part of rule consequentialism (as Mulgan does) or take the view that rule consequentialism evaluating our principles by referencing different worlds is necessary to avoid collapse (as Podgorski does) than you will think of me as being an act consequentialist, specifically an *indirect* act consequentialist.

As we discussed in the last chapter, that isn't wholly wrong. There is nothing that I say in this dissertation that cannot be said (or, indeed, *has* not been said (Moore G. , 1903) (Hare, 1981)) by a sufficiently sophisticated indirect consequentialist. But I believe SOC is better than indirect consequentialism for two main reasons. Firstly, it is less vulnerable to objections that derive from our reasons for actions (see Chapter 12, p. 241). Secondly, SOC makes the divide between orders very distinct and, as discussed in the previous chapter, we *ought* to keep this line sharp rather than fuzzy. As a result, even if SOC might be equivalent to indirect consequentialism in terms of what it says we actually *do,* I think SOC is the better theory overall, which is why I am not worried about collapse. I resist being characterized as an act consequentialist because I think doing so diminishes how important I think this division between orders is, and how critical it is to my overall project.

I also feel a lot of affinity for rule consequentialism. Like Hooker (or Brandt), I care a lot about the idea that we should internalize and adopt our first order principles and think that the cost of doing so is a large part of the reason why SOC often *doesn't* recommend act consequentialism at the first order level. But I am not a rule consequentialist, because unlike Hooker (Hooker, 2002, pp. 100-102) I am not committed to rules in any general sense, only

when doing so is recommended by SOC. This means I am untroubled by the example of situations, even frequent situations, where SOC does recommend act consequentialism at the first order level. But even in such situations, there is meaning in calling myself a second order consequentialist because being a consequentialist at the second order level is a *different thing* than being one at the first order level even if the former can lead to the latter. I am not a rule consequentialist or an act consequentialists, although I sometimes recommend acting according to rules, and sometimes recommend simple act consequentialism. I am a second order consequentialist, and everything else I believe about morality flows from that basic commitment.

PART 2: DEMANDINGNESS

Part 1 was about laying out the foundations of my theory, while the rest of the dissertation will be about applying it to various ethical problems. This part is about the problem of Demandingness. In chapter 6, I will lay out the problem and why anyone developing any kind of consequentialist theory needs to grapple with it. In Chapter 7 I will go over some consequentialist answers to the problem of Demandingness and why I think they don't work. In chapter 8 I will give my own answer to the objection

Chapter 6: The Demandingness Objection

Introduction

One of the most common and pressing problems that consequentialists face is the problem of demandingness. Simply put, most versions of consequentialism – and this is true to an extent even of non-act, non-maximizing consequentialisms (Case, 2016) – are highly demanding in a way that strikes most people as being unreasonable. There are a great many objections related to consequentialism's supposed over-demandingness, but they can be divided into two broad categories. First is the simple argument that no moral theory can be as demanding as consequentialism supposedly is. Secondly, there are more complex arguments that consequentialism demands that we sacrifice something uniquely and *especially* valuable, such as integrity or the separateness of persons, that should not be sacrificed or abandoned under any circumstances. I find both of these arguments to some extent unpersuasive. The first argument is usually defanged by consequentialists via deflationary theories of our intuitions, such as those I discussed in the second chapter, and while as I said there I think such theories can go too far, in this case I think they are right. The second, it seems to me, is not a special case, but a simple example of the larger argument between consequentialism and nonconsequentialism, and does not provide any *additional* considerations in the case of demandingness. In this chapter, I will go over the basic structure of the Demandingness Objection, including the arguments against it that are independent of consequentialism itself. Finally, I will provide a reason why we might nonetheless want to alter consequentialism to account for or defang the Demandingness Objection.

The Demandingness Objection

The basic Demandingness Objection is often brought out with examples. I will use Tim Mulgan's (2001, p. 4) fairly standard example of the Affluent. The Affluent is, as the name might suggest, a person who is fairly well off, and has the basic needs of life covered. In addition, he is someone who has the freedom to give some of his money away, and he does so regularly. He does not dedicate his entire life to the service of others, but participates when he can in protests, votes for positive social change, takes part in blood donations, and so on. On some particular day where he is free, he chooses to buy pricey theater tickets and have an afternoon of enjoyment. The problem arises that it would seem that most versions of consequentialism, classical utilitarianism most obviously, would condemn him for this action. After all, he clearly had other available options with far better consequences, namely sending that money to charity. But by the judgement of most, Affluent has done nothing wrong by enjoying an afternoon at the theater. If *all* he did was spend his money profligately that might be worthy of condemnation, but taking the occasional afternoon of for himself seems perfectly acceptable. Condemning him given the circumstances is downright unreasonable. This is the basic Demandingness objection: consequentialism (at least in its maximizing form) makes actions obligatory that seem to us clearly optional, and condemns us for courses of action that do not seem to be wrong.

One can also look at it another way, and say that the problem with consequentialist theories is that they do not seem to allow for the existence of actions that are less than ideal but permitted; this criticism is often phrased[23] as the claim that consequentialism lacks *options*. A

[23] e.g. Kagan (1998)

naïve reading of utilitarianism and other forms of simple consequentialism would say that there is only one permissible action in every circumstance, and that is the one that has the best possible consequences. There may be a rare occasion where multiple actions fit that criterion, but overall this is still a very limited set of acceptable options, whereas our commonsense morality and lived intuitions tell us that in most circumstances there is a very wide range of actions that are morally permissible, and that certainly it is far too demanding to allow for one to only do the best possible actions at all times.

A related criticism is that consequentialism also does not leave space for the inverse: actions that go beyond the call of duty. Consequentialist theories can thus also be criticized for lacking a category of *supererogatory* actions, actions that are morally praiseworthy but not obligatory[24]. That the classes of options and supererogatory actions do not seem to exist in a consequentialist morality seems a failing of the theory independent of the question of *what* actions are obligatory or supererogatory or permissible. One can admit either that contingent empirical facts might mean that consequentialism does agree with our intuitions on how demanding moral theories can be or that our intuitions are unreliable, while still considering it a weakness that simple consequentialism apparently cannot even in *principle* allow for supererogatory actions.

Let us examine this criticism in more detail to see why it is so compelling. Those arguing in favor of options might say that the problem is that it is virtually impossible, or at least extremely difficult, for a human being to follow consequentialist morality to the letter. This would mean, in turn, that even someone who is trying to hew as closely as can reasonably be

[24] e.g. Scheffler (1982)

expected to the demands of such a moral system is nevertheless acting immorally, in the sense that he or she rarely if ever makes what consequentialism supposedly claims is the only morally correct choice. This, says the critic, is a bitter pill to swallow: we would want to *praise* those who are being as moral as it is reasonable to expect them to be, not condemn them because they are acting unethically. Indeed, we would not even want to say that they are acting unethically at all, yet consequentialism – or at least the version of it being criticized here – would seem to say that even those people whom we consider the most moral to ever exist have never acted morally. Any moral theory with such an absurd consequence must surely be wrong.

This is a second order argument, an argument based on what we want moral systems to be and what we want them to accomplish. Further, it is an argument that can be made purely consequentialist in nature and is therefore very hard for a consequentialist to dismiss. After all, if a moral system claims that no-one who ever lived acted in a moral fashion, then most people are simply going to regard it as unreasonable and discard it. In addition, it seems extremely plausible that condemning people who act close but not quite optimally is very counterproductive, as it would create a sense of moral futility in both the condemned and others. And so on. It is not hard to come up with second order consequentialist reasons to worry about Demandingness, even if one is skeptical of the basic intuition that a moral system can 'demand too much'. Nonetheless, it is by rejecting that basic intuition that many consequentialists defend themselves against this objection.

Denialism and our intuitions

The most straightforward and perhaps most common consequentialist defense against the charge of over-demandingness is to simply deny that the relevant intuitions have any force (Singer, 1993) (Kagan S. , 1989) (Unger, 1996); let us call this strategy Denialism. I discussed

these theories at length in chapter 2 and, as I said there, many consequentialists go too far when it comes to being skeptical of our intuitions. From very legitimate worries about the accuracy of some of our intuitions, they extrapolate too much to an overall anti-intuitionism that is both implausible and very vulnerable to a Reid-style counterargument that undermining all our intuitions makes us unable to make any sorts of moral claims at all. This is especially true because the sources of knowledge the consequentialist would rather rely on, such as reasoning, are subject to precisely the same sorts of debunking arguments that they would use against our intuitions. A realistic theory of our intuitions, whatever it ultimately turns out to be, cannot involve a wholesale dismissal of all our intuitions. Having said all that, however, those who defend our intuitions against these charges, such as Mulgan (2001), are the ones who overstep when they think that blocking the Denialist argument in the general case necessarily means it is ineffective in this specific instance. Mulgan advances a general argument against wholesale skepticism of our intuitions that I think is largely successful, as I already discussed in chapter 2, but that argument does not necessarily license disregarding Denialism in the case of demandingness.

This is because a better Denialist strategy is to be *selectively* skeptical of our intuitions. For instance, we care greatly for close relatives because of kin selection, much more than we care for strangers. The naive Denialist would say that *all* intuitions that are the result of kin selection are suspect - kin selection merely happens because genes that code for it are more likely to spread, and why should we care what unthinking natural processes say about morals? But a more sophisticated account of these intuitions might say we are such creatures that require care throughout our childhood, and for whom close personal relationships are essential. Such relationships are a constitutive part of being a healthy, flourishing human, and necessarily come

with special obligations. This argument acknowledges the evolutionary reality - we are such creatures as the *result* of our evolution - but the constitutive argument is a normative one, that justifies the existence of special obligations towards close friends and relatives. It argues that our intuitions are correct in leading us to care strongly for our close friends and family, but may well nonetheless be wrong or limited in other aspects, for instance in causing us to *over*-privilege close relations above what is actually warranted.

One does not have to take the constitutive route either, as there are many different ways one might give a partially (as opposed to totally) deflationary account of our intuitions. For a moral naturalist, for instance, the analogy with our intuitions about the physical world becomes even more explicit: our intuitions tend to be on the right track because there are natural facts about the world that constrain the evolution of our intuitions, but these tend to have limits on their accuracy because evolution leads to local and not global optimums. And there are various other strategies one might take, depending on one's metaethical sensibilities. I am myself a naturalist, but the constitutive account also seems plausible to me, as do some other non-cognitivist accounts. These accounts have in common that they do not recommend a wholesale dismissal of all our moral intuitions, which is good because such a dismissal would be vulnerable to a Reid/Mulgan style counterargument.

Yet what all these accounts *also* have in common is a recognition that our intuitions have limits and are flawed. Moreover, they are flawed in some particular systematic fashions. For instance, our intuitions tend to over-correct. This is in fact true of our intuitions generally, not just our moral ones, and makes sense once one understands the evolutionary pressures involved. For instance, our pattern recognition skills over-correct because it is better to jump at a hundred shadows than miss a single lion. Similarly, our kin selection created instincts trigger for anyone

we interact closely with and not just kin because it is enough for evolution that most of the time close relations *are* kin, which is why birds will care for cuckoo eggs sneaked into their nests. Another systematic flaw is that because the evolutionary pressures of individual selection are stronger than those of group selection, our intuitions are likely to overprivilege the self over the group. Similarly, because reciprocal altruism is only evolutionarily advantageous in small groups where there is reasonable expectation of reciprocity, intuitions that result from it are likely to overprivilege the in-group over the outgroup, and to overestimate the moral importance of distance and familiarity.

Let's put it another way: an account of our intuitions that validates all of them as being totally accurate is even more implausible than one which rejects them wholesale. But our best understanding of where our moral intuitions come from and how they develop tells us that if our intuitions *are* wrong, then it is *more likely* that they go wrong in certain ways than in others. It is more likely that our intuitions lead us towards being more selfish than we ought to be than that they lead us to be less. It is more likely that our intuitions underestimate the moral obligations that we owe to strangers than that they overestimate them. It is more likely that our intuitions overestimate the moral importance of distance than that they underestimate them. And it much more likely that they think that morality is less demanding than it should be than the reverse. This is what I mean when I say that the intuitions that lead to the demandingness objection are uniquely suspect even if our more general moral intuitions are not. Even in a more moderate deflationary account that does its best not to throw the baby out with the bathwater, these sorts of intuitions are clearly the bathwater. This is why I find the Demandingness objection to consequentialism to be one of the weakest ones, and why I am suspicious of arguments that seek to justify it. I find it hard to not extend that suspicion to any argument that seeks to prevent

morality from being over-demanding, even as I myself have produced arguments that seek to defend consequentialism against the demandingness objection in a non-extremist fashion.

The most pressing problem the Denialist argument poses for people worried about the Demandingness objection, though, is that it tells us we have very good reasons to be worried about our sense of reasonableness when it comes to morality's demands. This is worrisome, because we often rely on precisely that sense to gauge the obligations a moral theory places on us. I do not want to pursue a purely Denialist strategy, largely because as I outlined earlier I think we have some purely consequentialist reasons to want our moral theory to not be over-demanding. But I am a Denialist to the extent that I think our intuitions about demandingness are highly suspect and possibly even should be thrown out or at least heavily revised. And the upshot of that is that we need a proper *theory* of demandingness, some sort of solid grounding for how demanding a theory ought to be. Such a theory would necessarily be a second order one.

The Integrity Objection and the separateness of persons

I will sum up briefly why I think that the Integrity objection or the separateness of persons argument do not produce additional considerations when we are discussing the Demandingness objection – which note is different from saying that these objections do not have any weight against consequentialism (or at least some forms of it). The Integrity objection is that consequentialism disrespects the *integrity* of persons because it demands that we set aside our personal welfare and projects to serve the general good (Williams, 1973). By demanding that we take the impersonal standpoint, we are left alienated from ourselves and unable to be fully human. Consequentialism thus disregards the importance of human integrity. As Mulgan discusses, the Integrity objection is closely related to the Demandingness objection, in that one of

the *ways* in which the demandingness of consequentialism is supposedly unreasonable is that it demands that we sacrifice our integrity.

But I have never been very impressed by that way of seeing the Integrity objection, as I imagine the poor farmer in the developing world to not be terribly impressed with it either. She might ask: what of *my* integrity, what of my projects? I must spend several hours and walk several kilometers every single day just to get water, that most basic of human needs. I must constantly set aside my own welfare for the welfare of my offspring. I must be ready to abandon my personal projects at any time, because I do not know how much time I *have*. If you care so much for people being able to pursue their personal projects and flourish, then you should wish for me to as well. Why must *I* give up my integrity, but not the affluent westerner?

But this is, in a sense, the entire consequentialist strategy. One argument often given in favor of consequentialism is that the most rational response to a value is to promote it. Far from disregarding the importance of integrity, the consequentialist might say that she is regarding it very highly indeed and being extremely demanding for precisely that reason. Just as she might ask you to give up some of your wellbeing so that a greater number might be better off, so she might ask that you sacrifice your personal integrity so that a greater number of people might regain theirs. It seems strange that the one arguing that consequentialism disrespects human integrity is the one advocating for a course of action that results in *fewer* humans having complete autonomous healthy lives. Of course, the entire overall argument between consequentialists and nonconsequentialists is in no small part about whether that is the right way to respond to things of value, but that is to underwrite that the integrity objection is not, or at least does not seem to me to be, a special case. This is merely the paradox of deontology recast in a different form.

The general problem is that consequentialism demands that we sacrifice things of value to increase the total amount of that value in the world - to make the world a better place. Bringing up the notion of human integrity as an important value does not seem to be to make the arguments in favor of the consequentialist *with regard to demandingness* either harder or easier. And I think similarly about many other ways of framing the demandingness objection, such as the separateness of persons argument. That is not to say that these objections to consequentialism are not interesting, or that consequentialists don't have to grapple with them and refute them if they wish to maintain the plausibility of their enterprise. Rather, it is to say that I think they have more to do with *other* objections to consequentialism than they do with demandingness as such.

There is perhaps another way of understanding objections such as the integrity or separateness of persons arguments: as objections to utilitarianism rather than consequentialism more generally. In fairness to the separateness of persons argument, both Rawls (1971, p. 37) and Nozick (1974) specifically target utilitarianism when they employ it. Rawls, in particular, is worried about the way utilitarianism views wellbeing as essentially transferable, a simple resource about which all we care about is that there is enough of it. Practically, this means that utilitarianism can condone extremely inequitable world states, the most extreme being Nozick's famous Happiness Monster where one individual has no diminishing returns from resources and therefore the greatest total utility is found by giving them all the world's resources. Conceptually, we don't actually care about wellbeing in the abstract, but rather we want *people* to be better off: part of the force behind the separateness of persons argument comes from the emphasizing that we care about wellbeing in the abstract *because* we care about people in the specific, and not the other way around.

In this form, however, the argument criticizes not consequentialism but rather utilitarianism's theory of the Good. A sufficiently modified consequentialism, that takes into account the importance of the distribution of utility and not just its total, and acknowledges that the wellbeing we care about is not some abstract transferable quantity but the wellbeing of specific existing people, seems to answer at least Rawls' worries about separateness of persons. On the other hand, I cannot see how such a modification would make the theory *less* demanding, and indeed it seems unquestionable that it makes it *more* so: in a consequentialist theory that truly made no distinction between persons, you could 'make up' for the lack of utility of the global poor by making the global rich sufficiently well-off[25].

Nozick's use of the argument is somewhat different and more extreme: he thinks that if the importance of the separateness of persons is to be taken seriously, then it cannot *ever* be justified to take from one individual's good to improve the lot of another. While Nozick is considering the legitimacy of a third party (the State) doing this, not the responsibilities an individual may themselves have towards others, it must still be noted that this view is just as extreme as the very demanding simple consequentialism, merely in the opposite direction. This makes it at least as untenable and indeed, if you share my worries about our intuitions in this matter, rather more so. I find such a morality simply implausible. At this point, the argument becomes less an argument about whether consequentialism is too demanding than an argument about the very idea of aggregating across multiple individuals at all, which leads us to the next point.

[25] Of course there are diminishing returns on things like money that make even simple utilitarianism demanding for the affluent, but the point is that introducing the importance of the separateness of persons as Rawls does surely cannot make it *less* demanding.

There is perhaps yet a third way of understanding the Integrity and separateness of persons objections: as *second* order objections rather than as first order ones. This is the argument that the consequentialist *mode of thought* is incompatible with substantial notions of personal agency or identity. This once again brings us away from specific arguments about demandingness to arguments about consequentialism more generally. As I said above, Nozick can be read to be saying something like this, as can Williams, and Korsgaard (1989) also argues looking at things purely from the impersonal point of view is incompatible with a robust notion of agency. Mulgan calls this the 'Transcendental Objection', and it calls back to earlier arguments I talked about with regards to partiality.

Like with Rawls, many aspects of this argument strike more against utilitarianism than against consequentialism more generally considered. Mulgan notes that if it can be shown that if a consequentialist theory can be formulated so as not to ignore a strong notion of agency, this objection is dissolved (Mulgan, 2001, p. 18). There have been attempts to modify consequentialism to do so, and we have discussed some of them already and will discuss more in the next chapter. What is relevant to this project is because this is a second order objection – that is, it is an argument that operates on the level of 'what sort of moral principles should we adopt' rather than 'what should we do' – a second order consequentialist like me can give a second order response. This needn't even be a very strange response. The Korsgaard piece cited is a response to Parfit, who argues against the importance of agency on the grounds of the non-identity of persons over time (Parfit, Reasons and Persons, 1987). Korsgaard gives several practical and pragmatic arguments as to why we should or need to care about agency in response, and many of these can be recast as second-order consequentialist arguments in favor of agency

(though Korsgaard would of course not wish to so cast her arguments). In this way, a second order consequentialist can provide pressure for our moral theories to be respectful of agency.

But to bring us back to the topic at hand, I do *not* think that a second order consequentialist can resist the charge of demandingness in this manner. Once again, this is in part because I believe they are largely separate issues. But more to the point, I think that as long as there is *any* pressure towards impartiality in our second-order considerations, taking human agency seriously entails a theory that is going to be very demanding. To give a clear example, I think of Kant's Humanity Formulation of the Categorical Imperative as being in fact very demanding: seriously making sure that in *none* of our actions are we treating people merely as means is a very strict criterion, and people are going to be uncomfortable with this in most of the same ways they are uncomfortable with utilitarianism's demandingness[26]. Kant tries to blunt this demandingness by introducing the idea of imperfect duties, but a duty is not actually less demanding because it is imperfect, merely less specific. To make our theory less demanding requires more that we *merely* imbue in it a respect for persons, as long as the theory also has at least some pressure towards respecting other persons as well. We can of course remove any of the latter, but such a moral theory would instead be implausibly *un*demanding.

A very similar thing can be said for the notion of Integrity, an argument best put forth by Elizabeth Ashford (2000). Ashford notes, first of all, that the kind of integrity we care about *can't* merely be an agents unified self-conception, as none of us are much impressed by, say, the white supremacist for whom the superiority of the white race forms a critical part of his self-identity. Saying that a moral system is 'too demanding' if it asks him to give that up is

[26] See also (van Ackeren & Sticker, 2014)

ridiculous, no matter how much of his personal projects are grounded in the overall white supremacist project and no matter how much changing that would alienate him from himself. Clearly, we care about a person's integrity only if that integrity is also grounded in the moral obligations that the person actually has, what Ashford calls *objective* integrity. But once we realize that the notion of integrity needs this kind of objectivity to be plausible, much of the force against utilitarianism dissolves, because it shows that utilitarianism is correct to sometimes allow the stringent demands of others to override our personal projects. One might think it does so 'too often', but that highlights that the problem we have is really finding the right level of demandingness, because claiming that a good moral theory will *never* obligate us to abandon our personal projects is just as implausible.

Ashford acknowledges that in the current state of the world, the demands of utilitarianism might be incompatible with maintaining one's personal projects – but she further argues that this is not a problem unique to utilitarianism. She shows how Williams' own moral commitments, if taken seriously, would demand that people sacrifice their personal projects, as would Scanlon's (1998) contractualist theory. I would also add that any plausible Kantian theory would do the same, and I believe in Ashford's conclusion: "in the current state of the world, any plausible moral theory has difficulty in showing how agents' impartial moral commitments and their personal commitments can be harmoniously integrated" (Ashford, 2000, p. 234). Once again, we see that the Demandingness objection is not a special problem for consequentialists. In fact, just like with separateness of persons, once we realize that the integrity we care about is *objective* integrity I do not see how incorporating it into our theory can possibly make it *less* demanding.

After all, we are now also asking people to not only sacrifice their personal commitments, but to change their self-conceptions to be more in line with the moral obligations they actually have[27].

In summary, even from a second order perspective, introducing the ideas of integrity or the separateness of persons bears little direct relevance to the demandingness problem. It introduces one more consideration that our moral theory needs to take into account, but it doesn't tell us anything about *how* our moral theory should balance being too demanding and not demanding enough. This brings us right back to my earlier point: it's not enough to simply list all the different considerations we need to balance against each other, we must also sketch out some means of actually doing such a balancing. It is not acceptable in this matter to simply appeal to some notion of intuitive reasonableness, because demandingness is a case where there are very good reasons to think our intuitions are highly suspect even if you are *not* generally anti-intuitionist.

Does the Demandingness Objection exist?

But following this line of thought introduces another, more extreme take: that even the basic Demandingness objection is actually just a well-disguised version of the fundamental difference between consequentialist and nonconsequentialist views. This position is articulated by David Sobel (2016, pp. 238-259), who calls it 'The Impotence of the Demandingness Objection'. Despite this rather eye-catching title, Sobel does not think that the objection has no force against consequentialism. Rather, he thinks that any force it does have comes from another, more fundamental objection to consequentialism rather than from the idea of demandingness

[27] Admittedly and also as with the case of separateness of persons this may not make a practical difference, as utilitarianism is still going to want people to give up on, say, white supremacist worldviews for instrumental reasons. Still, it is certainly not *less* demanding.

itself. Therefore, he says, we should reject consequentialism independently of the Demandingness objection or not at all.

The easiest way to consider Sobel's point is to use his own example: imagine a situation with two people, Joe and Sally. Joe has two healthy kidneys but only needs one, Sally has no functional kidneys. From the point of view of most consequentialist theories, Joe has a moral obligation to donate his kidney to Sally, but to many people doing so comes at such a high cost to Joe that he is not under a moral obligation to do so. Let us not quibble over the example and assume it is refined such that the latter is indeed the intuitive conclusion. When we say a consequentialist theory is too demanding in this instance, we mean it is too demanding on Joe.

But now consider things from *Sally's* point of view. She might well say that a moral theory that does *not* give Joe a moral obligation to donate his kidney to her is too demanding on *her*. You might instinctively think that this is not what we mean when we say a moral theory is too demanding, that we might instead describe Sally's complaint as saying that the theory is not demanding enough. But Sobel claims this is to beg the question: if we accept that a theory is demanding if it requires a high cost from those who follow it but not if it *allows* a high cost to those unaided by its followers, then we've already presupposed the distinction between requiring and allowing (which is *not* the same as the classic doing-allowing distinction, of which more below). But if we accept that there is a morally significant distinction between the costs a theory requires and the costs it permits, we've already – or so Sobel thinks – made the required break with consequentialism that the Demandingness objection is supposed to motivate us towards. Therefore, he argues that the Demandingness objection is not itself significant, but rather any force it has comes from this underlying distinction.

I think Sobel's way of viewing the objection is quite distinctive and enlightening, but that he is rather mistaken about the significance of it. It brings out, I think, that a theory that is undemanding on Joe in turn must ask a lot of people like Sally. You might see some similarities with how I addressed the Integrity objection earlier. Nonetheless, I don't think this is quite the same thing, and I don't agree with Sobel's take that this means that the Demandingness objection is merely piggybacking on some more fundamental objection to consequentialism. I think Sobel makes two different misunderstandings here.

The first misunderstanding is not distinguishing between moral theories in their *evaluative* modes and moral theories in their *action guiding* modes. Sobel's requiring/allowing distinction is not the same as the doing/allowing distinction we normally think of. The latter is a distinction that arises when we are deciding on courses of action, with some thinking that the difference has moral significance and others disagreeing. Let us sidestep that argument for the moment, though, because Sobel's requiring/allowing distinction arises when we are comparing *moral theories*, with Sobel alleging that consequentialist moral theories do not see a morally significant difference in the costs a moral theory accepts for the unaided and the ones it imposes via its moral obligations, while nonconsequentialist theories do. However, I don't think that this is actually what underlies the Demandingness objection.

The reason it seems strange to say that the nonconsequentialist theory is 'demanding' too much of Sally isn't because of some prior acceptance that costs the theory imposes matter more than costs the theory allows. Rather, it is because Sally is not the one in the situation that is choosing between options and therefore not the one whose actions are being evaluated as being obligatory, permissible, blame or praiseworthy, or similar. As McElwee (2017) puts it in his response to Sobel, in this situation Joe is the agent while Sally is the patient. Joe is the one being

asked by the moral theory to impose a cost on *himself*, on his own will. He is also the one in this situation who can be held responsible for his actions or failure to act. It's not like a theory that is too 'demanding' on Sally by Sobel's formulation is going to *condemn* Sally for failing to live up to her 'obligations' of allowing Joe to keep both his kidneys. Joe is the one with the choice here, it's his actions that are being judged. It is with reference to actions, not world states, that a theory is demanding or undemanding. When we ask how demanding our theories should be, what we are asking is to what standard our actions should be held – I point I will come back to extensively later.

Sobel's second misunderstanding has to do with the nature of objections themselves. Sobel argues that the Demandingness objection is 'impotent' because any actual force the objection has comes from the underlying intuition that a moral theory is worse if it imposes harm as a cost to complying with it compared to allowing harm to be done, size of harm being constant. But if we've accepted that intuition, he says, we've already rejected consequentialism, which means we ought to do so independently of the Demandingness objection or not at all. But the point of arguments like the Demandingness objection in the first place is to bring out these kinds of intuitions, to allow us to reflect deeper on our moral theories and what we want out of them. That doesn't make them impotent, it is them doing their job. Further, this supposedly underlying intuition does not actually strike me as a deeper one than the original phrasing of the demandingness intuitions, because I also don't agree with Sobel that the requiring/allowing distinction is the significant difference between consequentialist and nonconsequentialist theories. It seems more to me that the requiring/allowing distinction is the Demandingness Objection placed into a different form, a *restatement* of the objection rather than a new, underlying objection.

Now, restatements of old arguments are not without value and this one is no exception. Viewing the same problem from a different angle can often yield new insights. For instance, looking at the cost of what a theory allows in addition to what it demands may motivate you to consider more demanding theories, or so it does for me. But there are benefits and drawbacks to each way of stating an objection. The drawbacks to Sobel's version have a lot to do with his odd formulation of it, which in turn has a lot to with his first misunderstanding. Sobel alleges that if we accept the requiring/allowing distinction we've already rejected consequentialism, but here I think he is mistaking traditional consequentialism's impartiality or impersonal viewpoint as being about its action guiding mode when it is about its evaluative mode. A theory that is impersonal in its *action guiding* mode is one that does not generate obligations for individuals but merely generates obligations *qua* obligations on, I suppose, persons in general. Now, I will actually explore ideas of group action and collective responsibility in a later chapter, but that's *not* usually what people mean when they say that consequentialism takes the impersonal or impartial view! I'm very consequentialist in my views, but I still wouldn't say that classical utilitarianism is 'less demanding' on Sally because it obligates Joe to hand over a kidney, because in this scenario no moral theory is demanding anything of Sally because she's not making the decision. Now, if a moral theory claimed that Sally would have to reject the kidney even if it was freely offered (not that I think any plausible moral theory would do that, but perhaps it is one that says that averting your assigned fate of death by medical means is immoral or something) then *that* theory would be very demanding for Sally, but that's because she now has an obligation placed on her by the theory that seems unreasonable to most. I disagree with him that "The way the Objection measures the demandingness of an ethical theory reflects rather than justifies being in the grip of key anti-Consequentialist conclusions." (*Ibid.* p. 238)

However, thinking about the requiring/allowing distinction is a quite fresh and interesting way of looking at the Demandingness objection, even if it is the same objection viewed from a different angle and not a deeper prior assumption as Sobel claims. In particular, just as I earlier brought up second order concerns as to why the Demandingness objection should bother even hardcore consequentialists, the requiring/allowing distinction gives nonconsequentialists a reason to reject or soften the force of the objection. After all, when viewing things from a second order level, you might well think that a moral theory that asks more of the poor than it does of the rich is one we should reject. Considering what a moral theory allows to exist as well as what it demands might change your mind about whether classical consequentialism is really unreasonable.

Conclusion

All of this is to bring out what I want to be the central takeaway from this chapter, which is that the Demandingness objection is more than just a problem for consequentialists where their theory sometimes asks the unreasonable of us. It is a far more general problem, which is that we don't want our theories to be too demanding but also don't want them to be too *un*demanding. This challenge – finding the right level of demandingness – is one that every moral theorist must face, consequentialist or not. Many traditional answers to the Demandingness objection, from consequentialists or otherwise, do not succeed in this second, in my view more important, goal. In the next chapter I will go over some of these traditional answers in detail, and explain why I think we ultimately need a second order theory to truly *resolve* the problem of Demandingness.

Chapter 7: Traditional Answers to the Demandingness Objection

Introduction

In this chapter I will discuss some common consequentialist responses to the Demandingness objection and what I think their shortcomings are. In the last chapter I explained the basic problem of Demandingness and outlined one common consequentialist strategy to answer it, namely denying that the relevant intuitions have any force. Let us call this strategy Denialism or Extremism, both names are used in the literature. Overall, I find this strategy quite compelling, because even though I am not generally a Denialist about our intuitions I do think that our intuitions surrounding demandingness are the least reliable of our intuitions by far. However, I also wish to explore if it is possible to provide a consequentialist answer to the demandingness problem that is still in line with our intuitions, if nothing else because I think we do have good second order consequentialist reasons to not want our theories to be too demanding independent of our pre-theoretical intuitions, as I talked about in the last chapter. In this chapter I will discuss some strategies consequentialists have adopted apart from Denialism, why I think they are incomplete, and why I think a more systematic response to the problem, a true second order response, is needed.

Satisficing consequentialism

If one rejects Denialism, then the most straightforward way to defend a version of consequentialism from Demandingness is to build an allowance for options into consequentialism by relaxing the strict requirement for only allowing the best possible course of action. In this approach, typically called *satisficing* consequentialism (Slote, 1984), rather than demanding that one should do the action with the best possible consequences, the theory says

there is some threshold of good consequences whereby any action that meets that threshold or is better is permissible, and all others are forbidden. Depending on where the bar is set, this can allow for a wide range of permissible actions. We can even then say that actions that exceed the threshold are worthy of additional praise – and the more praise the more they exceed the threshold – thus allowing for supererogatory actions. This variant of consequentialism is straightforward, simply building the allowance for options into the base of the theory, and thus is very good at achieving its aim.

Satisficing consequentialism is arguably no less elegant and no more complex than maximizing consequentialism, since in principle all it changes is the location of the threshold. However, it does have additional problems of justification that maximizing views don't have. Maximizing views are straightforward to justify: it is easy to argue that the best action is the one with the best consequences and also that one should always strive to do the best action. Satisficing consequentialism is *also* easy to justify, in the sense that it is easy to motivate the mere existence of a threshold lower than the maximum. The Demandingness Objection itself serves as such motivation. However it is extremely difficult to motivate any *particular* threshold in a way that does not seem arbitrary.

For instance, imagine you are pursuing a course of action whose consequences are just above the threshold of satisficing consequentialism. Because any course of action with more negative consequences would drop you below the threshold, it would be condemned – no matter how slight the difference is. In other words, you have two courses of action where one is condemned but the other is acceptable, even though the difference between them might be very small – one involves slightly more littering than the other, for instance – which just seems arbitrary and weird. Similarly, if you are just below the threshold, *any* positive change to an

action that pushes you even slightly into the 'good' range would seem to be far *more* positive in this theory than our intuition says it actually is. As we will talk about in more detail later in Chapter 11, a lot of the problems that arise around satisficing consequentialism are very similar to those that arise around threshold deontology, which is about an entirely different kind of threshold altogether. In both cases, the problem is one of *arbitrariness*: what makes *this particular* threshold the right one? Even if we establish a vague range instead of a precise threshold, that does not change the fact that the theory does not give us a good reason to place the range there as opposed to somewhere else (Ellis, 1992) (Alexander, 2018). Furthermore, regardless of where we place the threshold, satisficing consequentialism in fact cannot simply be maximizing consequentialism with a lowered threshold, but must be more complex than that. Simply lowering the threshold creates additional problems that do not exist for maximizing consequentialism.

For example, in *Against Satisficing Consequentialism* (2006) Ben Bradley notes that if the total utility is already above the level demanded by satisficing consequentialism, then actions that *lower* utility, even murder, seem to be morally acceptable provided utility remains above the critical threshold after such actions. To be clear, this is very different from the normal case where consequentialism might be criticized as being too accepting of things like murder. Consequentialists might say that murder is permissible or even obligatory if it is the only method to save lives, but in this scenario murder is *not* having overall good consequences, it is simply having not bad *enough* consequences to dip below the threshold. But surely, even in such circumstances actions like murder should be forbidden (again, absent other considerations like the murder itself saving lives or being in self-defense). On the flip side, straightforward satisficing would seem to imply that doing nothing but sitting in your chair is praiseworthy if it

happens to be true that the consequences of doing nothing are far above the threshold, which also seems somewhat absurd (Chappell, 2019). Mulgan (2001, pp. 128-139) comes up with a very similar set of objections, showing through thought experiments that one can come up with situations with satisficing (or, as he calls it, sub-maximizing) consequentialism where one is permitted to take actions, including murder, that lead to worse consequences even when a better action is clearly available and costs nothing.

Satisficing consequentialism can be refined to avoid these problems. An example is Richard Yetter Chappell's 'Willpower Satisficing', which abandons the straightforward threshold in favor of saying that an action is permissible if its consequences are at least as good as any other action the agent could perform with a cost to that agent of up to some amount X (Chappell, 2019). However, such sophistication comes at the price of being, well, more sophisticated and thus a loss in simplicity. More importantly, while it might allow the theory to escape from some of the criticisms of Bradley and Mulgan above, it does *not* offer escape from the demand for justification of the particular threshold (in this case, the particular cost X). Chappell recognizes this (*Ibid.* pp. 11-12) and in fact strives to provide such a justification. He fleshes out the notion of 'cost' by reference to a theory of blameworthiness; this is in fact similar to my own approach, though our theories of blameworthiness are very different[28]. Without such a fleshing out, though, no version of satisficing consequentialism, however sophisticated, can be complete. The lesson to take from this is not that satisficing consequentialism is inadequate to answer the problem of Demandingness – indeed as I just mentioned my own theory can be seen as a form of satisficing at the first order level – but that it creates an *additional* demand for

[28] My own theory of blameworthiness is the subject of the next few chapters

justification over maximizing consequentialism. This means that it is incomplete without a second order theory of some kind.[29]

To return to Mulgan (Mulgan, 2001), he does make a distinction which is helpful here and that is between *strategic* sub-maximization and actual sub-maximization. As we have discussed earlier at many points, even act consequentialists will often say that it is usually better to follow proven rules than to sit down and calculate the utility of every single action one takes. That is because the time and effort taken to calculate each and every course of action can be better spent elsewhere, and the opportunity cost means that it is overall usually better to follow simple rules. Similarly, even a maximizing consequentialist would say that as a *practical* matter one should generally aim for the best action one can manage, without wasting too much time trying to work out what is the best *possible* action in each and every circumstance. This is still maximization however, not satisficing – you are still trying to do the action with the best consequences, it is just that *when you factor in decision cost* wasting time trying to figure out the actual best action in all circumstances has worse consequences than doing the action that you can reasonably determine is the best given your limited information and the time you can spare. This sort of strategic sub-maximization is something even a maximizing consequentialist can embrace without trouble and leads to none of the problems that actual satisficing consequentialism runs into. You will never be able to set up situations where it is permissible to do something even when a better option is clearly available, unlike with what Mulgan calls *blatant* sub-maximization, because strategic sub-maximization says you should do the most clearly available best option, and is only sub-maximizing because that isn't always the *actually* best option (but it

[29] As I argued in chapter 4, I also believe we have independent reasons to want a second order theory. However, non-maximizing forms of consequentialism make the need for one even more acute.

is often enough that it's not worth spending the extra time and effort making sure). Mulgan and I both agree, however, that strategic sub-maximization alone cannot solve the problem of demandingness, because even just aiming for the best reasonably available (as opposed to best *possible*) action is actually a very demanding criterion.

And indeed, even *blatant* sub-maximization can be argued to be too demanding, as Spencer Case does in a recent paper (2016). If we look at the Demandingness objection as a question of whether the theory allows for options or supererogatory actions, than satisficing consequentialism is a solution. But Case says that the real problem is whether the actions the theory *gives us the most reason to do* are extremely burdensome or not. When you put it in *those* terms, Case says, satisficing consequentialism is less demanding than maximizing consequentialism only with the addition of some other assumptions that alter what the theory says we have most reason to do – assumptions that must necessarily be very controversial and dubious.

One could note, furthermore, (though Case does not) that if looked at through this lens, Demandingness becomes a problem for more than just satisficing consequentialism. That the actions the theory says we have the most reason to do are extremely burdensome is a problem for many nonconsequentialist theories as well. This touches back on a point I mentioned earlier and will mention again: Demandingness is not in fact a special problem for consequentialist theories, though it is often seen that way, but rather a more general issue that all moral theories have to struggle with. To be more specific, maximizing consequentialism does have the particular problem that it does not seem to allow for options and supererogatory actions, which is something most theories can do. But any moral theory that gives us reason to care for others and asks that we sometimes put their good ahead of our own – which is, of course, the vast majority

of them – will also say that the actions that we have the most reason to do are ones which we would find extremely burdensome.

Agent-centeredness

Satisficing consequentialism is not the only way that consequentialists have tried to solve the problem of options. The other main method is to introduce some sort of agent centeredness: some way of adjusting the consequentialist calculus to allow for more morally acceptable options by increasing the weight of the concerns of the moral agents themselves. The classic example is Scheffler's Hybrid Consequentialism (1982), which modifies consequentialism with an 'agent-centered prerogative' that allows one to give more weight to their personal projects. It can also be done with more complicated schema, such as Douglas Portmore's Dual Ranking Act Consequentialism (2008). Portmore's theory is fairly involved, but briefly it states that an action is permissible if there is no other action that generates *both* more total wellbeing for others *and* more total wellbeing overall, counting while adding an additional weight to the agent. From the perspective of the problem of demandingness, however, only the second part is important (the first is there to accommodate intuitions that it is more acceptable to sacrifice your own wellbeing for the overall good than that of other people), and it is a straightforward agent-weighting (Portmore even calls it Schefflerian Utilitarianism). It thus has all the problems of agent weighting, with the most problematic being that any proposed weighting, like any proposed satisficing threshold, seems both arbitrary and, if precisely defined, clearly wrong.

Although the mechanics of agent weighting are quite different from satisficing consequentialism, the problems it has are rather similar to the latter, including that one can construct scenarios where the theory allows for bad actions even when there are clearly available low-cost alternatives. Kagan (1984) notes that Schefflerian Utilitarianism seems to permit

causing harm for the sake of fulfilling one's personal projects, and not just allowing it. Scheffler seems loath to embrace the doing-allowing distinction wholesale, seeing it as abandoning the consequentialist project. In later papers (Scheffler, 1992) tries to give pragmatic (or, as he says, 'quasi-practical') reasons for why we should have our moral theory permit people to allow harm to focus on their personal projects but not cause it without relying on nonconsequentialist considerations. However, as Mulgan points out (2001), there are only two possibilities here. Either these reasons are good enough to apply to all cases, in which case it would seem that we ought to embrace a general restriction on causing harm in pursuit of impersonal goods as well, which Scheffler certainly does not want. Or they aren't, in which case one can still construct cases where we seem permitted to cause harm in pursuit of our personal projects, which Scheffler *also* does not want. Of course, if we do embrace a doing-allowing distinction wholesale it could allow us to escape the Demandingness objection, but at that point it is that distinction rather than agent-centeredness that is doing the work.

Agent-centeredness also does not escape from the Case-style critique from earlier, or at least only escapes from it by embracing a very uncomfortable consequence. Obviously, Schefflerian consequentialism says that we are *permitted* to depart from promoting the general good to pay attention to our personal projects, and isn't simply a modification of maximizing consequentialism with an agent centered weight. That latter would imply that we are *obligated* to *not* donate to charity if doing so would infringe on our personal projects too much, which seems to me blatantly absurd. But if so, it still seems that the actions we have more reason to do are very burdensome. Furthermore, it once again seems like the agent-centeredness isn't doing the work, but that rather it is the fact that we are permitted to not maximize the general good for *any* reason that is what allows us to escape the Demandingness objection – that is, it is simply

another form of satisficing consequentialism. It would still be a worthwhile addition if agent-centeredness gave us a *reason* to not maximize the good, and this clearly is what Scheffler and others are going for, but that approach is not without its own problems.

This is because the real problem with agent-weighting is the one we outlined in the previous chapter. The intuitions that lead to agent-weighting are, simply put, extremely suspect. Indeed, they are perhaps the most suspect of all our intuitions. This is ultimately the real substance of Kagan's 1984 critique. He notes that Scheffler fails to give us sufficient reason to think that we ought to allow agent-relativism in the way he describes. It's undeniable that we do as a matter of fact care about our personal projects and place our interests above that of the general good, but it's quite another thing to say that we are justified in doing so. As Kagan puts it: "Personal independence may constitute an implicit appeal for agent-centered prerogatives – but what is the rationale for *granting* this appeal? (Surely not the mere fact that the appeal is made.)" It's not so much that such an rationale cannot be found, but that Scheffler does not provide one. Kagan offers a few rationales of his own, such as a constitutive argument from the nature of humans, but any of these would need to be greatly expanded on for the theory to not collapse into implausible egoism or parochialism. In short, for agent-centeredness to be a viable theory, it is not sufficient to simply develop it as a first order theory – just as with satisficing consequentialism, it requires a substantial second order theory underlying it to be viable.

Collective Consequentialism

We discussed Collective Consequentialism (CC) in Chapter 5, and why one of the reasons Rule Consequentialism is often seen as synonymous with it is that it is the primary means by which rule consequentialists address the Demandingness Objection (Hooker, 2002, pp. 160-175) (Mulgan, 2001, pp. 67-103). CC mitigates the problem of being overly demanding by

distributing its demands onto the population as a whole. In its simplest form, it says that we each only have an obligation to do as much as *would* be enough to obtain the best consequences were everyone (or most people, depending on the theorists view of compliance) to do that much.

This formulation has several problems, the first of which is the one we have discussed already in Chapter 5. Since Collective Consequentialist solutions to Demandingness make reference to other possible worlds to formulate their required levels of obligation they are vulnerable to the distant world/ideal world objection. More sophisticated formulations avoid this problem only by adding enough specificity (i.e by adding mechanisms to deal with non-compliance) that they come to resemble simple act consequentialism in terms of demandingness (and also in other ways, but it is the demandingness that matters most for our discussion). After all, if the theory is modified so as to be responsive to noncompliance with the ideal principle, then in situations where there is widespread non-compliance with that principle – *such as the world we live in* – the demands on those who do follow it will be very high. (Mulgan, 2001, pp. 87-89).

Now, I argued in Chapter 5 that this sort of 'collapse' into act consequentialism in some hypothetical scenarios need not be a problem for SOC, and I stand by that claim. But it *is* a problem if you want to argue that CC is *inherently* less demanding than simple act consequentialism, as it can be shown to not be in several plausible scenarios (*Ibid.).* There are several reasons why I do not take this tack of avoiding the Demandingness Objection by direct appeal to SOC formulated in Collective form, and one of them is that I agree with Mulgan that it does not by itself provide a satisfactory answer to the objection. Even CC, at least if it is to be plausible and to avoid some of the stronger forms of IWO, still seems likely to be very

demanding. The simple version of CC, by contrast, is not only vulnerable to IWO but I view it as implausibly *un*demanding.

In other words, just like previously, Collective Consequentialism alone does not address the question of exactly how demanding our theories ought to be in a satisfactory manner. It seems at first as though it might give us a promising way forward, because the question of "how much would each of us have to do for it to be true that, if everyone did that much, the world would be best" at least seems non-arbitrary. This is a step up from, for instance, basic satisficing, which does not give us a way to decide upon where the threshold is without further elaboration. However, the problem of arbitrariness creeps right back in when we start asking what we are obligated to do in situations of non-compliance (for example, how much are we obliged to 'make up' for others not following the principle?), bringing us right back to where we started. Avoiding that by simply saying we have *no* obligation to make up for the lack of others not only throws us into the teeth of the IWO but is also, in my view, flatly implausible on the face of it.

Mulgan does have another objection to Collective Consequentialism, though one which I find much less compelling: the Wrong Facts Objection. The crux of this objection is that under CC how much each person has to sacrifice is dependent on many factors that Mulgan thinks shouldn't be relevant. For example, straightforwardly if there were half as many people in need then we each would have to sacrifice half as much, and ten times as much if there were ten times as many people in need. Mulgan thinks these extreme differences seem unreasonable given the different situations:

> it seems ridiculous for Affluent to donate only one-fifth of 1 per cent of her income, simply because there are *only* 1 million people starving, rather than 50 million. Similarly, it seems unreasonable to demand that the [Affluent in the case of Many Poor] give up all of her income simply because there are 2,500 million people starving rather than *only* 50 million. (*Ibid.* p. 91)

I find the first example more compelling than the latter one, but neither seems to me to be a serious problem. After all, it seems reasonable for the demands on each of us to be in *some degree* dependent on how many people in the world are actually in need. What seems unreasonable here is the strict linear relationship between the two factors – but it does not seem to me that CC *must* propose a linear relationship here. If everyone adopted principles that caused our societal and economic systems *as a whole* to be more equitable, I don't think we'd each have to sacrifice exactly twice as much to help 500 million as 250 million. The world isn't that linear. Mulgan says that "Rule Consequentialism falls into these mistakes because it presupposes a rigid relationship between the sacrifice required of any particular individual and certain features of their global situation" (*Ibid.* p. 93) but I'm not convinced it actually does.

The other kind of Wrong Facts objection Mulgan raises, however, takes aim at precisely the point that the world is more complex than that. In order to effectively decide how much Affluent *does* need to donate, she must seemingly know which economic model is the correct one to determine where best to donate, she needs to know exactly what outcomes her money will have, etc. To be clear, she needs to know this because the assumption of CC is that each of 'us' – where 'us' means those with the means and responsibility to give, and how that is determined is a whole question in itself – need to give as much as would be sufficient to solve the ills of the world were all of 'us' to give that much. This means that how much Affluent needs to give is dependent on factors such as how many people are in need, how many of 'us' there are, and how much money is actually needed if spent in the best way (and what that way even is). All of this is information it is unreasonable for Affluent to have, and it's especially unreasonable for her to need all this information to simply decide whether she needs to give to give 10% or 15% of her income to charity.

Again, though, I think this is a rather unfair view of what a properly elaborated CC would actually ask us to do. Perhaps rather than saying we should each give X% of our income to charity it asks that each of us do X% of the *work* of transforming our world into a more equitable one. In such a case, I do not think the relationship between the amount of work we have to do and these other factors is as rigid as Mulgan believes. And as he says: "I do not claim that these empirical differences should have no effect at all on Affluent's obligations. I claim only that they should not affect her obligations to the extreme extent that Rule Consequentialism implies." (*Ibid.* p. 94) It's certainly true that many Collective Consequentialists talk mainly in terms of giving X% to charity but we should, in the spirit of charity ourselves, think of that as a **simplified model** used to get across the basic idea of CC rather than what a plausible Collective Consequentialism, once it has been wholly elaborated, would actually say.

Mulgan does have a point that the obligations on each individual seem to be very dependent on facts it is unreasonable to expect individuals to have knowledge of and therefore it is impossible for individuals to work out what their obligations are under CC but, ironically, this is a problem that arises from thinking of the theory *too individualistically*. As I discussed in Chapter 5, philosophy is inherently a collective enterprise and when I am doing second order ethics I should be taking in the views of, and taking advantage of the knowledge of, other people. One of the reasons to draw a distinction between the two levels in the first place was to say that second order ethics is the kind that we do when we have the ability and leisure to consult with experts and sort through complex information, to decide on those principles that guide us when we do *not* have that leisure. I think it's kind of ironic that Mulgan argues against a moral theory that distributes the work of solving the problems of the world onto the collective instead of individuals by assuming that the work of determining *how* that work is distributed still falls on

the individual. It is unreasonable to expect me to need to be an expert on global finance, economics, politics and so on in order to decide how much I need to give to charity, but it's not unreasonable to expect me to listen to the experts when I am formulating those moral principles I will live by. *Any* moral theory will involve some level of this kind of division of labor and deferral to experts, CC is not alone in this.

What this also brings up is something else I discussed in Chapter 5, which is that there are many ways of understanding 'collective responsibility'. I do not believe in Collective Consequentialism, by which I mean the view that we each are only obliged to do as much as *would be* enough if all of us were to do it. That does not mean I do not believe in the idea of collective responsibility or collective action at all. Chapter 10 will be a long exploration of exactly that topic. I will save the discussion why the kind of collectivization I believe in is different from CC to that chapter. What I will say here is that I agree with Mulgan that it is not at all clear that embracing an idea of collective responsibility leads to less demanding theories *in itself*. In addition, CC is very vulnerable to IWO and similar objections. What I do think about collective responsibility is that incorporating it into our moral theory has substantial effects in what *kinds* of actions the theory recommends we take (as we'll talk about in Chapter 10), but not nearly as much in the *level of obligation* we have. In short, I do not believe that collective notions of responsibility are a good solution to the *Demandingness Objection*, which is why I will not discuss them here further.

Mulgan and pluralist theories

In *The Demands of Consequentialism* (2001), Tim Mulgan gives an overview of many consequentialist theories and how they fare in regards to demandingness. He views simple consequentialism, the basic idea that the only right thing to do is that action which has the best

consequences, as unreasonably demanding. He then considers many other consequentialist theories, including rule consequentialism, satisficing consequentialism, and agent relativism. He considers all of these approaches to be inadequate in some way. I have discussed several of the arguments he uses already and will discuss some of his specific arguments later, as some of them might be said to apply to my own theory as well, but much of the rest of it would be going too much off topic. What I will say is that my overall criticism is that he gives too little weight, as I discussed earlier, to intuition-debunking arguments, and that this leads to a deep weakness in his final theory.

Mulgan's own approach is to distinguish between two different kinds of moral relationships (he calls them two moral realms) that generate different kinds of moral commitments: the relationship (he calls the Realm of Necessity) that we have with all beings of moral value and the one (the Realm of Reciprocity) that we have with moral agents that we share an equal and mutual relationship with. These different kinds of moral realms generate different moral imperatives and then he devotes the rest of the book to explaining how to manage these imperatives against each other and what to do when they conflict, finally coming up with a theory that he thinks is not too demanding which he calls Combined Consequentialism. In this schema, one adopts simple consequentialism as the best theory of the Realm of Necessity and rule consequentialism as the best theory of the Realm of Reciprocity, then uses a variant of hybrid consequentialism to balance the two.

But this way of dividing up moral realms or different kinds of moral relationships triggers all my worries about intuitions that reinforce our inborn parochialism and why they are inherently suspicious. These kinds of arguments - arguments that say that there is a morally relevant difference between our moral duties to members of our community and our moral duties

to members outside of it - are arguments that are both worryingly apt to being used to support abhorrent positions and worryingly dependent on intuitions that are highly suspect. To some extent this is a strawman: using Mulgan's idea of two moral realms to support racism would not be straightforward, and would not bear much relationship to the ways racism manifests in real life. The problem is that if you are making these kinds of arguments to justify our intuition that morality cannot be too demanding, it strengthens the Extremists because they can note that the fact arguments of a similar *form* can be used to support racism or (perhaps more worryingly) nationalism, and other forms of troubling parochialism. This in turn gives us more reason to doubt the validity of both the argument form and of the intuitions used to justify it. Given that on most plausible accounts of intuitions those that arise from or support our ingroup-bias are, morally speaking, highly suspect in the first place… well, this is what I mean when I say I find the Extremist position hardest to respond to.

Mulgan's two realms are more about distinguishing between two different kinds of moral reasons than about two different kinds of relationships, and he takes pains to say that the boundaries between the realms are fluid and can overlap. But his approach is still susceptible to the same sort of worries, and is particularly so because to a large extent it is constructed backwards. That is, Mulgan notes that how demanding his Combined Consequentialism is will depend on what sorts of weighting we give to the two different kinds of moral reasons, and to complete his theory we would want to give a weighting that accords to our reflective intuitions. This sort of theory may produce an account that is both consequentialist and in accord with our intuitions (as was Mulgan's aim), but only because it has been deliberately constructed to be so. This feels rather inadequate to me, because it does not itself provide a means to interrogate or justify those intuitions; a lack which is particularly worrying because, again, in this matter our

intuitions are especially suspect. Thus, because it is all too easy to adjust his Combined Consequentialism to be more or less demanding, it is hard to see how he can ensure it is the right *level* of demanding – for surely it is at least as important, if not more so, that our morality does not demand too little as that it not demand too much. Mulgan recognizes this, of course, and recognizes that his theory needs to be tuned appropriately, but fails to actually point to a way by which such a tuning may be accomplished.

Now as mentioned above I have my own response to the demandingness objection. But my approach is not to refute consequentialism's demandingness but rather to *defang* it. I argue that making moral judgements on actions and persons is something that should itself be evaluated on consequentialist grounds, and that when we do so we find that it is wrong to condemn someone who does not give all their money to charity. But in a lot of ways my position is still an Extremist position. I think that when you fail to give most of your money to charity you are doing something forgivable, something human, and that it would be not merely inadvisable but *unjust* to condemn you or take you to task for it or judge you for it or hold you to account, but that there is still *some level* on which what you are doing is wrong. I think that niggling guilt you should have for doing something at least nominally wrong is valuable - and we are on something like third-order consequentialism now I fear - in the same way that our niggling suspicion of intuitions that conform to or support our ingroup bias is valuable. I think that no reasonable moral system can *condemn* us for being human, but also that a moral system is inadequate if it doesn't exhort us - however gently - to be as good as we can possibly be.

Interestingly, when Mulgan *does* give reasons for adjusting the demandingness of his theories or the other candidate theories he discusses, those reasons are what I think of as 'second-order' reasons. This is because he thinks that rule consequentialism is what should give reasons

in the Realm of Reciprocity, and rule consequentialism is very much a second order theory in my framing, especially as he describes it. However, Mulgan is of the opinion that rule consequentialism, and by extension these sorts of second-order considerations, cannot be used to balance these two realms (p. 228). Instead, Mulgan relies on a version of Scheffler's agent-centered prerogative, though his is more of a community centered one. The problem is not that Mulgan fails to provide a precise method of balancing his two realms – such would be the project of several lifetimes, not merely a single book – but that he doesn't sketch out a *method* which one might use to approach the problem of balancing the two realms that is not dependent merely on our (in this case highly suspect) reflective intuitions. And this criticism, as I noted above, is one that extends back to Scheffler as well.

In summary, Mulgan's approach is a pluralist one and has the standard strengths and weaknesses of pluralist approaches. It combines the strength of all its component parts: in Mulgan's case, he thinks that in the Realm of Reciprocity rule consequentialism is the best theory, while in the Realm of Necessity simple consequentialism is the best one. Thus, his account gets to have the best features of both, while avoiding their weaknesses. But pluralist theories must also provide a means to weigh their different components against each other and resolve conflicts between them, and here is where I think Mulgan's theory falls down. Now, to a certain extent Mulgan achieves what he sets out to do: his aim in this book is to provide a consequentialist theory that *can* be reasonably but not too demanding. In this he largely succeeds. But what I want to do is sketch out a consequentialist approach that isn't just capable of being only reasonably demanding but which sketches out what 'reasonable' demandingness might look like, in consequentialist terms. My approach too, is sometimes a pluralist one (as we

will see in the discussion of thresholds), but it retains, by virtue of its second-order nature, an underlying unified logic that allows it to escape the standard weakness of pluralist theories.[30]

Beyond consequentialism

I have mentioned a few times now that I don't think that Demandingness is a special problem for consequentialism, but a more general problem for moral theories of any type. To be precise, however, there are some particular aspects of the objection that work more against consequentialism than other theories. It is true, for instance, that consequentialism has more trouble making room for options and supererogatory actions than nonconsequentialist theories do. But any moral theory that tells us we sometimes need to care about the impersonal good, which is most of them, and any moral theory which says we sometimes need to put that good ahead of our own, which is all the plausible ones, must answer the questions of *when* and *how much*. If we look at things like Case does, it seems that the actions that most moral theories give us the most reason to do are very burdensome, so there is a sense in which they are all very demanding. Of course, those moral theories may say it is permissible to take other actions, but maximizing consequentialism may equally say that we should not punish those who take less than the best action unless that would in itself have good consequences. This is a response one can see as early as Mill (1861) and my own theory can be said to be a developed version of this response. Either way it underlies that maximizing consequentialism is neither uniquely susceptible to this problem nor uniquely unable to resist it.

On the flip side, while it is easy to adjust our moral theories to be less demanding, it is hard to do so while *also* ensuring we do not make them too undemanding. We do not want our

[30] Though, as it happens, I do not agree with Mulgan that one must appeal to pluralism to resolve the demandingness objection

moral theories to justify selfishness and egoism, nor parochialism and excessive partiality. Moral theories must also provide sufficiently good reasons to be less than maximally demanding, as simply appealing to our intuitions in this matter is not an available option as it might otherwise be. This is where more traditional answers to the Demandingness objection from non-maximizing consequentialists fall short, but it is also where several ordinary moral theories fall short. Every moral system needs to have a developed theory of Demandingness and an answer to these sorts of questions, and merely being nonconsequentialist or non-maximizing does not actually allow one to escape this requirement.

Fortunately, philosophers are coming alive in recent years to the idea that Demandingness is a problem for more than just consequentialists. Kantians especially are mindful of the fact that their theory can be very demanding and try to find their own answers; see for instance Walla (2015) and Ackeren & Sticker (2015). Others, like Stephen Harris (2015) discuss Demandingness outside of Western philosophy; Harris in particular constructs an Indian Buddhist response to the charge of over demandingness. Others argue against the idea can be too demanding outside of the traditional Denialist responses, like Goodin (2009) and Berkey (2014). Many of the ideas these authors bring up I have already discussed and will discuss in later chapters. In particular, Walla argues that a proper response to the demandingness problem requires that our moral theories call for the creation of just institutions and not just private action, and Berkey argues that climate change produces a special problem for those who claim that morality can be too demanding and should cause us to further doubt out anti-demandingness intuitions. Both of these ideas I will discuss at length in Chapter 10, which is about collective responsibility.

Conclusion

The takeaway from this chapter should not be that I think that the answers consequentialists have given to the Demandingness objection I discussed in this chapter are entirely unpromising or innately flawed. As I said in the beginning, I don't really believe in singular knockdown arguments that can allow us to rule out entire types of theory. In fact, I think that many of these responses can potentially answer the objection given sufficient development, and my own theory can be seen as a developed form of satisficing consequentialism. However, such development needs to be done. In essence, I find them largely incomplete rather than wrong, because in the absence of a strong theoretical foundation that can tell us *how* demanding a theory should be and engage with and criticize our intuitions we cannot answer the real problem of demandingness. Part of the reason for this incompleteness is that the traditional discourse around Demandingness has overfocused on a relatively small aspect of the problem – that maximizing consequentialism seems overdemanding and leaves no room for options – and missed the larger, much more pressing issue of demandingness. Namely, that many moral theories fail to provide a way to gauge what the right level of demandingness is. In the next chapter I will develop my own second order consequentialist response to the problem of Demandingness, as well as press the argument that some sort of second order theory, consequentialist or not, is needed to truly solve it.

Chapter 8: Second Order Consequentialism and Demandingness

Introduction

In the last chapter I contended that any solution to the problem of Demandingness fails or is at least incomplete without a second order moral theory. Partly this is because Demandingness is itself a second order objection, and so demands a similar level response. But it is also because any purely first order theory, regardless of whether it is consequentialist or not, is going to run into the same problem: attempting to balance concerns of over and under-demandingness without a framework to judge what that might mean. This task is made additionally problematic because our moral intuitions, regardless of how trustworthy they may otherwise be, are unusually suspect in this case. That doesn't mean that one must be a second order consequentialist, of course, and for addressing the problem of demandingness I could see several other plausible routes one might take. I won't, however, discuss them in much detail, since my primary objective here is to advance a consequentialist second-order theory.

I will argue that if we take consequentialism seriously on the second order level, it follows that our moral judgements must themselves be justified by a consequentialist calculus. This means that as long as we have second order consequentialist reasons for our moral judgements to not be too demanding – and I have already argued for this – then we have a means to address the Demandingness objection. Much more importantly, however, we will have a framework for judging exactly how demanding our first order theory should be, one grounded in underlying moral principles and not just intuitive judgements we have good reason to doubt. While this approach is unusual and engenders its own objections, I think it is very promising. In

this chapter I hope to convince you that this approach is at least initially viable, while the next chapter will be devoted to answering those objections.

Consequentialism and demandingness

It is commonly considered obvious that both options and supererogatory actions cannot exist in a classically utilitarian or similarly maximizing consequentialist moral systems, but I disagree. I believe that this conclusion is based on the unstated assumption that any course of action that is not the one(s) that the moral system prescribes is by definition immoral. But this assumption is open to question, especially in the case of consequentialist systems of morality. It seems obvious because that is the case for most nonconsequentialist moral theories. In most rule-based systems of morality, the emphasis lies on defining those actions that are immoral – the ones that result in one of the rules being broken. Thus any action that does not abide by those rules *is* by definition immoral, since the concern of such a system of morality is to define what is immoral.

A consequentialist system, however, does not provide a heuristic for avoiding immoral courses of action but one for determining morally optimal courses of action: the best courses of action are the ones that have the best consequences. This means that it is no longer necessarily true that any other suboptimal course of action is immoral. A consequentialist system of morality is not forced to define the option that has the best consequences as the *only* morally correct one. One can alternatively argue that all consequentialism really says is that the *best* action – the one we have most reason to do – is the one with the best consequences (Norcross A. , 2006) (McElwee, 2010). But it does not simply logically follow from the basic definition of consequentialism that we must regard all other actions as immoral, but rather only as worse than that best action, i.e. *suboptimal*.

It might be argued that consequentialists *should* believe all other actions are immoral, but this would actually be an unhelpfully rigid definition of morality (McElwee, 2011). In fact, this rigidity would actually rule out supererogatory actions entirely. Surely we have normative reasons to do supererogatory actions – if we didn't, they wouldn't be praiseworthy. But that means we have *more reasons* to do supererogatory actions as opposed to merely obligatory ones. This is simply another way of saying that the supererogatory course of action is better than the merely obligatory one. Now, the consequentialist that takes this tack does owe us an explanation of precisely the difference between supererogatory and obligatory actions if they wish to allow for the former. More generally, they owe us an explanation of how we should regard suboptimal actions if we are not going to regard them as immoral. It is at this level that I think satisficing consequentialisms fail and where I think we need a second order theory. But that the best action is the only morally acceptable one is an additional claim on top of the basic theory that consequentialists are by no means obliged to make.

Some consequentialists go so far as to discard notions of moral or immoral actions entirely. This position is *scalar consequentialism*, most famously defended by Alistair Norcross (2006). Norcross argues that utilitarianism cannot give an account of *right* or *wrong*, only *better* or *worse*. The only thing there is to say about a course of action is how good it is relative to other available courses of action. "There is no further fact," he says, "of the form 'x is right,' 'x is to-be-done,' or 'x is demanded by morality.'"(*Ibid*., 228). Further, Norcross argues that there is no conceptual connection between wrongness and censure or blame. Utilitarians will blame or censure people based on the utility of doing so, but Norcross rejects hashing out wrongness in terms of this kind of blame. Doing so is impossible, he claims, because our concept of wrongness is constrained by certain principles that this definition would violate. These are:

1) If action x is wrong, then an action y done by someone in exactly similar circumstances, with the same intention and the same consequences, is also wrong. (Norcross calls this 'universalizability');

2) if someone has done the best they can do, and does very well indeed, then they have done nothing wrong. (*Ibid*. 225-6).

But as McElwee (2010) points out, the very same things could be said, with equal plausibility, of *blameworthiness*. If action *x* is blameworthy, action *y*, done by someone in exactly similar circumstances, with the same intention and the same consequences, is surely blameworthy too. It is just as implausible to censure someone for doing the best they can as it is to say that they are wrong – this is the very core of the Demandingness objection! Norcross rejects defining wrongness in terms of a utilitarian notion of blameworthiness, saying that "it is absurd to say that [one] has done something wrong just in virtue of the fact that it is appropriate or optimific to punish [them]" (p. 256). This is a perfectly reasonable claim – save that it is, surely, *exactly as absurd* to punish someone that has done nothing wrong! Conversely, if the utilitarian is, as Norcross claims, comfortable with censuring, blaming and even punishing an innocent person if it would be optimific to do so (*Ibid.*), why should they suddenly balk merely at also claiming they were wrong? It is not *per se* inconsistent to be willing to give up on these principles in the case of blameworthiness but not in the case of wrongness, but it seems a strange and unmotivated place to draw the line. If you care about principles like universalizability, it seems that you should care about them in both cases and so be equally unwilling to embrace a consequentialist definition of blameworthiness.

This very argument, which I call the *conceptual* objection, is one we will discuss in some detail the next chapter because I in fact do want to define blameworthiness on the basis of the consequences of holding someone blameworthy and so must defend myself from it. But I agree with McElwee that Norcross is drawing a distinction without a difference here. If Norcross is right that consequentialists cannot define 'right' and 'wrong' in terms of the consequences of doing so because it would disregard key features of wrongness like universalizability, then those same arguments apply equally well to praise and blame. Conversely, if one can show (as I attempt to in the next chapter) that a consequentialist notion of blameworthiness can *preserve* these principles (at least in all but very distant possible worlds), then that defense can be equally well applied to a consequentialist notion of rightness and wrongness because the latter concepts are constrained by *the same principles*. What we cannot have is what Norcross seems to have in mind in this paper: giving up on the notions of 'right/wrong' and 'obligatory/permissive', but remaining justified (on consequentialist grounds) in praising or blaming others. If we give up on the first we should also give up on the second. But to move to second order concerns for a moment, we *do* want our moral theory to justify praise and blame, and to tell us when each is merited: if a moral theory cannot do this, that is a reason to reject it. So both Norcross's approach and a hypothetical one that bites both bullets seem unviable here.

And yet, I somewhat agree with Norcross, in that I do think that first order consequentialism does not have the ability to say more about a course of action than whether it is better or worse than others. But I think *second* order consequentialism does have the means to define blameworthiness and praiseworthiness and – indeed, *pace* Norcross, *therefore* – right and wrong. In fact, to step away from consequentialism for a moment, I think this is the business of second order theories in the first place, and that no first order theory of any kind is well equipped

to define right and wrong, for reasons well discussed by now: in the absence of a second order theory the only way to do so is by reference to intuitions we have good reason to doubt. In the next section I will describe the consequentialist notion of blame and praise that Norcross seems to have in mind, and after that I will go into the second order version. In the next chapter I will defend that version from the criticisms Norcross and McElwee raise here, but I need to lay out my theory first before I defend it.

The consequences of judgements

Allow for the moment that a form of consequentialism that is by most people's standards too demanding is indeed the correct moral theory. Now imagine someone who adheres to these standards as closely as he or she reasonably can, allowing for such things as human weakness and irrationality. By the consequentialist definition this person is acting suboptimally. But should they be *condemned* for doing so? Only if condemning them would lead to a better outcome than not doing so. For most of the classically immoral actions – murder, theft and so on – it is plainly the case that this is true (at least in unexceptional circumstances). But for the crime of not being perfect, it is plausible that judging the person to be acting immorally is unlikely to do any good and far more likely to do harm.

Some of the reasons have already been stated: if we are deemed bad people for not living up to an unreachable ideal, it seems likely that many of us would become disgusted with the moral system that does so and deny it; this is exactly where the opposition towards consequentialism comes from in this particular case. Indeed, a person who does the best action they reasonably can should be praised for doing so, even if it is not the best possible action available to them. Thus there exist actions which we cannot censure people for not doing and are

morally praiseworthy to perform, even though they are strictly suboptimal in consequentialist terms; for example, giving a significant amount to charity but not as much as you could.

In a recent paper, Rick Morris (2017) advances the same argument: that consequentialists distinguish the judgement of an action from the judgement of praise or blame for that action, and also that if certain plausible claims about human psychology hold true that they are in fact *morally required* to not blame people for feeling to meet certain very demanding obligations, or even to praise people for performing actions that fall short of the best possible action. It may seem strange, that a person should be praised for doing seemingly 'immoral' things, but there is no actual paradox here, despite appearances. The confusion only arises because we assume that there should be a connection between which courses of action are deemed morally suboptimal and the standards by which we judge whether or not people are immoral. But as discussed above, consequentialists need embrace no such strict connection.

Understanding this distinction also blunts some other criticisms of simple consequentialism's demandingness, such as those that imply it to be self-contradictory (Cullity, 2004). The argument goes that we cannot be required to aid people in achieving non-legitimate interests – that is, we cannot be obliged to help others do something immoral. But those who live non-altruistic lives are, according to the extremely demanding form of simple consequentialism, acting immorally. Thus, extreme demandingness either requires us to aid those with illegitimate interests, or it does not require us to aid people who live non-altruistic lives, making it too *limited* in its beneficence.

This argument actually makes several related mistakes. It is not a contradiction for consequentialism if the most moral action leads to other people being immoral, as long as that *was* still the best possible action. This argument imports a non-consequentialist principle, rather

than showing a contradiction. Cullity would argue that it might be true that he is doing this, but that merely highlights that the simple consequentialism is even more nonintuitive than it first appeared, since it seems committed to saying that we should aid people in fulfilling non-legitimate interests, and a moral system cannot be morally compelling if it asks us to aid others in fulfilling interests that same moral system condemns. But as I've been saying, the consequentialist criterion of right action doesn't map onto judgments of people, nor interests, in such a one to one fashion. It is a mistake to think that a consequentialist is obliged to say that one who acts suboptimally is automatically immoral, or that an interest that is anything less than perfectly altruistic is illegitimate. Consequentialism distinguishes between not taking the most moral action and being an immoral person. Or rather – because putting it that way makes it sound like consequentialism has an *additional* feature when it in fact lacks one – consequentialism does not have a strict connection between those two things.

There is an opening for a counter-counter argument here: while there may not be a strict logical necessity that a person who does an action other than the best one be deemed immoral, it *should* be the case that this is so. That is, a moral system which does not have this concordance between the moral status of actions and of the people who do those actions is a deficient one – a second order argument. But this opens up further second order considerations: is it actually *true* that this is something we want of moral systems? In practice, concordance is only perfect for an unrealistic moral system that most of us would deem to be unacceptably strict. Most of us would wish our moral system to have some room for forgiveness, some allowance for people to make mistakes and the occasional bad call. But any system of morality that does have this room, and thus any reasonable system, has an imperfect concordance between the judgment of people and

of actions. Thus, lack of perfect concordance cannot be an effective criticism of this form of consequentialism.

And for the most part, consequentialism also has concordance, both between the judgments of actions and of people and between those judgments of people and our intuitions. In particular, not sacrificing one's own well-being is forgivable, where forgivable here means that it is not moral to censure someone for it. Indeed, forgivable is not quite the right word: it would be *immoral* (unforgivably immoral) to censure someone for a reasonable level of imperfection; it is an open question what exactly that reasonable level is, or whether it would always lead to better consequences to condemn that person. But that is beside the point; I only wish to make the point that consequentialism *can* have its own category of actions that are similar enough to options that our intuitions about judgments concerning whether or not people are immoral are mostly satisfied: courses of action that are immoral, but forgivable in this extremely strong sense of that word.

The defender of options may not be very impressed with this line of argument making room for options in consequentialism. They might admit that someone whose actions were suboptimal might not be blameworthy in a consequentialist system. However, they might say their intuition is not merely that such a person cannot be censured, but that he is not guilty of any wrongdoing at all. Yet I would wish to claim that consequentialism is capable of concurring with our intuitions even in that regard, and to do so I will turn to second order arguments.

Second order consequentialism about judgements

This response to Demandingness need not be thought of as a second order argument; Morris does not present it as one, for example. I think it helps, however, to put this argument in a larger second-order framework, in which all parts of our moral system, not just how we assign praise or blame, are put through a consequentialist analysis. By seeing it like this, we can more clearly compare it to the closest equivalent, satisficing consequentialism, and why it has an easier time with the latter's problem with arbitrariness. Again, a similar arbitrariness worry arises with Mulgan's account, or Portmore's DRAC, or with threshold deontology, or in other places, and in all of these cases second-order consequentialism can provide us a framework that is capable of addressing these concerns. That doesn't necessarily mean the first-order theory is doing no work – perhaps something like Dual-ranking is necessary to accommodate the permissibility of self-sacrifice, for instance – but it does mean that the first-order theory alone is incomplete without a second-order framework to place it in.

By being explicit about the difference between orders, and drawing a sharp divide, we can be clear that what we are *not* saying is that suboptimal actions should not be labeled as immoral even though they 'really' are. It is a criticism sometimes raised about consequentialism that it can be deceptive in this sense: for example, it can be argued that a consequentialist should pretend to be a deontologist because that will lead to better results. I will talk about this 'self-effacing' criticism later[31], but this is not what I am suggesting here. Instead, I am arguing that the *definition* of a moral or immoral action depends on the consequences of labeling it as such, in addition to the consequences of the action itself. That is, an action is immoral by definition not

[31] Chapter 11 (p. 201)

just because it leads to negative consequences but also if condemning it, or those who perform it, leads to better consequences than not doing so. An action is obligatory, again by definition, if it has positive consequences and condemning someone for not acting in such a way leads to better consequences than not condemning them. An action is supererogatory if it is not obligatory but praising someone for doing it would lead to greater well-being, and so on. It is not that we should consider these actions as such but rather, I submit, that second-order consequentialism in fact requires us to *define* these actions as such – because doing so produces the greatest good.

A few points of clarification are in order here. This does not require there to be a person who actually makes that judgment, any more than a person needs to judge if an act is optimal in order to make it so. The act is optimal by the nature of its consequences, not by virtue of someone declaring it so. Similarly, an act is immoral if censuring it leads to better consequences, regardless of whether or not it actually is condemned. Further, there is room for discussion – though it is beyond the scope of this chapter, as it relies on many empirical facts and is indeed the work of a lifetime – about how much weight should be given to the consequences of the action itself and how much to the consequences of judging it. As long as some weight is given to the latter, however, my argument shall hold.

There is also the specter here of a vicious infinite regress: Must we also decide if it is moral to label an action as moral, and so on ad infinitum? Upon further reflection, however, this problem remains but a specter. From the beginning, it remains true that the correct thing to do in any situation is to perform the action with the best consequences. It might not be the only moral action, but it is unquestionably *a* morally permissible action. Any action that leads to the best consequences is similarly guaranteed to be moral. Labeling an action as moral or immoral, obligatory or supererogatory, depending on which label would have the best consequences, is

similarly guaranteed to be a moral thing to do, thus halting the regress. Admittedly, doing so may be complicated, but that is always true of act consequentialism, and just like with actions, the solution is to go by rules of thumb for the most part, only deciding on an individual basis when there are extraordinary circumstances and time to spare. Essentially, the theory has actual (or as Mulgan calls it, blatant) sub-maximization at the first-order level, but only *strategic* sub-maximization at the second-order level. Thus it does not require a third order theory to justify how suboptimal the second order theory can be, etcetera, because the second order level is actually maximizing.

To some extent, the form of consequentialism I am here advocating bears similarities to satisficing theories of consequentialism. But it would not be entirely accurate to characterize it as such. In satisficing consequentialist theories, the existence of options is built into the moral system from its foundations and therefore supererogatory actions exist in it independently of any other facts about it. By contrast, in my suggested second order consequentialism – where moral and immoral actions are defined as such depending on whether or not doing so would have better consequences, in addition to the consequences of the acts themselves – the existence of supererogatory actions is dependent upon contingent facts. It is possible, perhaps probable, that designating some actions as supererogatory instead of obligatory might lead to better consequences in the long run, but it is also possible, I submit, to imagine a world (that consists of a species very different from humanity, perhaps) in which the best outcome is had by not defining any actions to be supererogatory. It might even be the case that this is so in *our* world, though I judge it highly unlikely.

This in turn means that my theory is immune to many criticisms of satisficing consequentialism. For example, it is fairly easy for me to argue that condemning certain actions

such as murder – provided that they are deliberate and avoidable – leads to better consequences *regardless* of how high the level of utility is beforehand. Similarly, satisficing consequentialism cannot distinguish between doing and allowing – that is, we want to say that actively lowering the threshold is immoral but merely standing by is acceptable, provided that the total utility remains above the level demanded in all cases – whereas it is plausible that my theory can distinguish between the two. Condemning the former unquestionably has good consequences in almost all circumstances. Punishing the latter at least arguably does not, since censuring someone for something that was *not* the direct result of their intent and actions will have negligible effect on their actions in the future and thus have few consequences of any sort, let alone good ones.

Furthermore, if supererogatory actions exist in a maximizing consequentialist framework, they do so because of human imperfection. That is, the reason regarding moral optimality as supererogatory rather than obligatory results in better consequences is that making it immoral to be merely imperfect would lead to few people accepting any theory that does so as a valid moral theory. And the reason that few people would accept such a theory is assuredly for the reasons described earlier when discussing why our moral intuitions in this matter may be compromised: people are selfish and lack perfect knowledge, rationality or discipline and thus expecting them to be perfect would be (immorally) unreasonable.

Thus it is still to some extent true that there is in principle no limit to the sacrifices that consequentialism asks of beings, provided that those being are unlimited in their capability to abide by the ultimate demands of consequentialism. If people are, and it is a legitimate empirical question as to whether they are not, then supererogatory actions may not exist even in this framework. However, it still provides the possibility of such actions existing, and if they do exist

then there is in principle a limit to the level of sacrifices consequentialism can make obligatory of us – a limit that cannot be raised unless human nature itself is changed, regardless of the circumstances. Wishing for there to be such a limit irrespective of *any* contingent facts whatsoever is still too much to ask of maximizing consequentialism.

In particular, for a race of ideal people – who are not selfish, have perfect knowledge, and so on – supererogatory actions do not exist at all, in contrast to satisficing theories of consequentialism or any moral system which produces moral options from some other principle than the one I have used here, such as saying that supererogatory actions are not obligatory because of the cost to oneself. Indeed, more generally, by my definitions the same action may be obligatory for one race, supererogatory for another and even perhaps forbidden for a third. In other words, the definitions of what is moral and immoral are relative to the moral capabilities – their psychologies and capacity to respond to moral motivation, but also the extent to which having the ability to live their own lives is necessary for their wellbeing etc. - of the beings involved.

This is not nearly as radical a thing to say as it may seem, since at least to some extent everyone does think that these definitions are relative by species. For example: when a male lion takes over a pride from another, one of the first things it does is murder all the children of the previous male. We would not wish, I think, to say that it acts immorally by doing so, because it is not reasonable to expect a lion to be bound by the principles of morality. Assume I am right in that a consequentialist should define an immoral action as such if doing so leads to better consequences than not doing so, not merely on the basis of the consequences of the action itself. If so, it should be fairly obvious that they should not call this action immoral because labeling this action as immoral (or moral, for that matter) does not lead to better consequences because it

is unlikely to have consequences at all. Thus it is pointless to so label those actions. Saying that an animal cannot act immorally hardly seems to be a controversial thing to say, thus indicating that our definitions of moral and immoral actions are already relativized to species.

It might be that defining an action as moral or immoral depending on the consequences of doing so might also lead to morality being relative to cultural factors. Unlike the species example earlier, this does not strike me as immediately obvious. At least in the case of whether or not there are supererogatory actions, it seems to be that if labeling some actions as supererogatory rather than obligatory leads to better consequences then it is because of fundamental and universal features of human psychology, not cultural factors. Nevertheless the moral system being advocated here may well lead to different standards of morality for different societies, though I do not think that is a weakness. After all, we in fact do alter our moral standards at least to some extent depending on the circumstances.

Again, none of this re-defining of what it means for an action or a person to be moral or immoral changes what is the *best* action, which is also always morally correct. For simple act consequentialism that is still the action that has the best consequences (for utilitarianism, that action that results in the greatest amount of well-being). Similarly, for SOC it does not change the fact that the action that conforms to the moral principles that lead to the best result if adopted is both morally correct and the best action. The only thing it changes is that this action may no longer be obligatory, nor necessarily the *only* possible action that is morally acceptable. Further, the above statement is true of all beings regardless of species or society. What is relative to the circumstances is *how* suboptimally you can act while still acting morally – i.e. in a manner not worthy of censure. For an animal, this can be very sub optimally indeed (one can easily argue

that an animal cannot ever act immorally precisely *because* it is never worthy of moral censure) whereas for a human it may be only a little sub optimally.

A brief aside – SOC does sometimes allow for situations where the 'most optimal act' in terms of consequences is actually wrong, as in the Transplant case we discussed in Chapter 5. However, that is because SOC is concerned with what would be the optimal set of principles on the second order level rather the optimal act at the first order level. However I do not believe that SOC can answer the Demandingness Objection in the same way as it addresses the Transplant case, since even the optimal set of principles is likely to be very demanding. SOC detaching the morally optimal thing to do from being the *only* morally acceptable thing to do is what allows it to address the Demandingness Objection, just as it is what would allow simple consequentialism to answer the objection. But what SOC thinks is the morally optimal thing to do is not going to always align with a simple act consequentialism.

It may seem strange, perhaps even contradictory, that the morally optimal thing to do is not the *only* moral thing to do. But recall my earlier argument that we have *more reasons* to perform supererogatory actions than obligatory ones. Far from being strange, this is in fact how we already think about supererogatory actions. Thus this apparent weakness is fact a concordance with our intuitions on the matter. Furthermore, I believe that our intuitions do indicate that we have a moral imperative to (as an example) give to charity. This imperative might be weak enough that doing so is not obligatory, but if there were no moral imperative whatsoever to do so then why would giving to charity be morally praiseworthy?

Thus, I would also argue that we do not equate the fact that there is a moral imperative to do something with that something being morally obligatory. A consequentialist is still bound to perform to the best of his ability only those actions that lead to the best possible consequences,

but doing so is not obligatory (because what it *means* for an action to be obligatory is that not doing it is worthy of condemnation and thus immoral), merely encouraged; i.e. morally praiseworthy. It seems to me that such a category of actions matches very closely to what our intuition demands of supererogatory actions. The major 'difference', if such it can be called, is the emphasis laid on the fact that while these actions are morally acceptable they are still suboptimal. This emphasis is not present to the same degree in deontological moral systems, but I do not see how it could be a weakness of a moral system that it asks of you to perform as perfectly as possible without denouncing you for your failure to do so.

Thinking about how to judge actions in this way doesn't just solve many problems with demandingness, it can also be applied to clear up issues in other cases where our intuitions can lead to troubling or contradictory results without some framework by which to judge them, that being in many ways the main benefit of this second-order approach. For instance, I have been silent in most of this dissertation about what theory of the good you might adopt, largely because I want to develop a general framework that is compatible with several different theories of the good. But let us suppose that, as I alluded to in the last section, one adopts a theory of the good in which agency is held as a thing of great value, in which case a second-order consequentialist would want to adopt some kind of agent-centrism into their first order theory, diverting it from simple consequentialism. This is a different approach to solving the demandingness objection than I focused on in this section, more akin to Mulgan. But it lacks the weakness of that approach that I elaborated on last section, because you have a framework to judge how *much* agent relativism to import into your theory, by focusing on the second-order consequences of doing so. As Kagan would put it, you have a rational for *granting* the appeal to agent-centrism beyond the mere fact that the appeal was made.

Conclusion

In summary, the argument that consequentialism and other similar forms of consequentialism cannot have space for options and supererogatory actions arises from the assumption, which I have hopefully shown to be erroneous, that consequentialism demands that any action that is not the morally optimal one must *necessarily* be considered immoral. Not only is this not necessarily true, but since it is certainly not guaranteed to lead to the best possible results a consequentialist arguably is morally bound to *not* hold this belief. Instead, we should let consequentialism define an action as moral or immoral depending on whether praising or censuring it would lead to the best results, not merely according to the results of the action itself. If that is the case, then an action is supererogatory if praising leads to good results but censuring those who do not perform that action leads to worse ones. It might be, as a matter of fact, that no such category of actions exists, but at least the possibility of their existence becomes feasible.

More importantly, using second order consequentialism as a framework for determining the criteria by which we judge actions and people solves the arbitrariness problem that would otherwise arise. Any other developed second order theory would do this as well, but SOC has the benefits of consequentialism's simple unifying principle, as well as a clear set of criteria to use for determining exactly how demanding our moral theory should be. Defending this theory of blameworthiness from the criticism that it is too unintuitive or that it does not capture important features of blameworthiness is the subject of the next chapter.

PART 3: BLAMEWORTHINESS

This part is a defense and explication of the theory of blameworthiness I introduced in the last chapter. Chapter 9 is a long defense of this theory of blameworthiness against many possible objections. In Chapter 10, I will apply my theory of blameworthiness to cases of group or collective responsibility.

Chapter 9: Defending My Theory of Blameworthiness

Introduction

In the last few chapters I discussed how second order consequentialists might answer the Demandingness Objection by adopting a theory of blameworthiness that says that we should hold some person or actions blameworthy in circumstances X if and only if adopting the principle that said person or actions were blameworthy in circumstances X would lead to the best consequences in the long run. I have thus far focused mostly on the benefits of adopting this theory of blameworthiness, and the reasons why I think it is worth seriously considering. However, it cannot be denied that this theory is quite revisionary, and asks us to reconsider some of our intuitions about blame and the rightness and wrongness of actions. In particular, a lot of what we might have thought to be inherent features of blameworthiness my theory takes to be merely external and contingent – only part of our account of blameworthiness because of their instrumental value. In this chapter, I will try to give an answer to these and other worries, and defend this theory of blameworthiness generally, for I do think that even despite these objections this is the best theory of blameworthiness. This will open the stage for using my account of blameworthiness fruitfully in other areas besides Demandingness, which will be the subject of the following chapter.

I will discuss what differentiates my account of blameworthiness from superficially similar instrumentalist theories of blame, and why many objections to the latter do not apply to my account. I will then try to defend my theory from those objections which I believe *do* apply to it, such as the worry that it can allow for extremely undesirable judgements – such as blaming the victim – in the right circumstances. I will argue that this is only true in circumstances so different from any in reality that it doesn't create a real problem for my theory. Assuming we are

dealing with human beings in a world that mostly resembles our own, I will argue that my account is revisionary but not outrageously so, and that where it *does* ask us to revise our judgements of blameworthiness it is correct to do so.

Second-order consequentialism and blameworthiness

My account says that we should judge people as praise or blameworthy on the basis of whether internalizing those judgements would lead to the best consequences or not. This seems intuitively nonobvious because it is bringing in concerns most don't think should apply to our judgements of blameworthiness. Surely whether or not an action is blameworthy depends solely on properties of the action itself, not second order externalities that might affect the consequences of blaming someone for it? Whether a person is good or bad should depend on the actions they actually *do*, and not on whether it would lead to good consequences to internalize the principles of judging them good or bad.

There are a few things I can say in response to that. Firstly, the objection here is somewhat overstated. I certainly do not believe that there is *no* connection between the rightness or wrongness of an action and whether or not it is blameworthy/permissible/obligatory etc. Indeed, my definition of those terms depends in large part on the consequences of the action itself, both directly and indirectly. To go back to our discussion of adoption and internalization from Chapter 5 (pp 73-74), adopting a theory of blameworthiness means believing in it, committing to it, and spreading it to others. The consequences of adopting such a theory are of course dependent in no small part on the consequences of the *actions* which the theory judges as good or bad. This is because those are the actions such a theory would encourage or discourage in oneself and others.

My theory is simply not dependent *solely* on the actions they take, but also on external factors – that is to say, context. When put that way, it maybe isn't actually such a strange way of defining blameworthiness. After all, we already make a distinction between things being 'wrong' and being 'blameworthy'. When we judge whether or not something is blameworthy, we do take into account the actions rightness or wrongness, but we *also* take into account many other factors that are related. For instance, we take into account whether or not the person was in their right mind, whether they were in possession of all relevant facts (and if not, we then have the further question of whether that ignorance was innocent or negligent), whether they were acting under duress, etc. For a consequentialist who defines an action's rightness or wrongness in terms of its consequences, all of these additional factors are separate from the question of whether the action is right or wrong (i.e. separate from the consequences of the action itself). The way to take them into account is precisely to focus on the long-term consequences of judging the action as blameworthy. We do not judge people acting under duress as being blameworthy for committing wrongful acts *because* it makes for far better consequences if we place the blame on the person who is forcing them to do things. And so on.

My position is that first-order consequentialism can only put actions on a ladder from best to worst, as with scalar utilitarianism (Norcross A. , 2006). Further, I would argue that for a consequentialist to be able to gauge which actions are permissible, impermissible, and obligatory, they would *require* a second order theory of the sort I have been discussing. My theory thus makes a distinction between an action that has less than the best consequences and one that is immoral, where the first order theory is used to judge one thing and the second order theory the latter. But I would like to argue that to a certain extent people already accept a distinction between the 'best' action and those actions that are permissible. That is, most people

are willing to accept that within the set of all the actions that are permissible, some are still better than others, while within the set of actions that are impermissible, some are still worse. If you think there is *nothing more to be said* about an action beyond whether it is permissible, impermissible, or obligatory, then this argument fails; but that is a pretty rare position. Most people still distinguish between actions that are better or worse, even if all the actions in question fall into the same category of permissible or impermissible.

I do not wish to overstate my case here: my account in unquestionably revisionary. Defining whether an action is obligatory or impermissible based on the consequences of internalizing the judgement of such is unusual even for consequentialists, who tend to base it on the consequences of the action itself (or, for rule consequentialists like Hooker, the consequences of internalizing the principles that allow for such actions). My theory defines blameworthiness in terms of when it is appropriate to blame, instead of the other way around. If adopting and internalizing the principle that a certain course of action should be blamed is part of that theory of blameworthiness for which it is true that adopting it would lead to the best consequences, then that is what it means for that course of action to be blameworthy.

The question of whether to define blameworthiness in terms of when it appropriate to blame or the other way around is by no means a settled one – both approaches are plausible. Consequentialists (Sidgwick, 1874) are more likely to determine appropriateness to blame first and define blameworthiness in relation to that. This is because consequentialists tend not to believe that actions are good or bad *simpliciter* (Norcross, 2006). A course of action (e.g. murder) might be generally bad because it almost always leads to bad consequences, but there will be exceptions to that general rule. As such, courses of action aren't inherently blameworthy, but rather are blameworthy if it is true that it would lead to the best consequences to blame the

actor. Whereas if you believe that certain courses of action are inherently morally forbidden then you might think that doing that action makes you blameworthy by definition, and you would define appropriateness to blame in terms of blameworthiness.

As a consequentialist it should be no surprise that I take the first approach rather than the latter, but complicating matters is that I take a multiple order approach. Thus, at the *second* order level, I first determine what theory of blameworthiness would lead to the best consequences if adopted and then define what is and isn't blameworthy on that basis and then internalize it. But that first order theory, once adopted and internalized, might well define some actions as inherently blameworthy. As discussed in Chapter 4, this is the point of the separation between orders – when deciding whether or not an action is blameworthy, I do so by calling on the first order theory I have internalized, not SOC.

By taking the approach I have, I have committed myself to some fairly unusual claims, and I judge some actions as blameworthy that many people might not, and similarly for permissibility. I wish only to point out that this approach is not *as* unusual as it might seem at first glance, since taking other factors into account besides the rightness or wrongness of the action itself is something we do all the time. This chapter is largely an exercise in what Enoch calls limiting the loss of plausibility points: I am not denying that my theory is unusual, but only trying to convince you that it is not nearly as troubling as it might initially appear. But to do that, it is critical to be clear about what my theory is, and what it is not. Mine is a theory of blameworthiness, but not a theory of blame, and in the next section I will discuss why that means some objections to my theory are in fact off-target.

Instrumentalism and the phenomenology of blame.

My account might be described as an *instrumentalist* account of blame, such as those of Smart (1961), Vargas (2008) (2013) or Jefferson (2019). For reasons we will discuss in this section, I think such a label would be misleading, but there are undeniably similarities. An instrumentalist account of blame claims that what we are doing when we make ascriptions of moral responsibility is attempting to achieve some end to which that ascription is an instrument (hence the name). Smart argues that the purpose of blaming others is to influence their future behavior in a positive direction, and many other instrumentalist theories follow at least somewhat in his footsteps by arguing that ascriptions of moral responsibility have the purpose of influencing people's actions or causing them to develop into more responsible agents. In principle, though, instrumentalist theories could have a number of different plausible ends for the practice of blaming, or a combination of several.

The comparison with consequentialist theories is obvious and naturally it has not gone unnoticed by instrumentalists. To quote Jefferson: "I take consequentialist considerations to be central to the justification of our practices of holding each other responsible. They go beyond a moral assessment of the quality of the action to a justification of our decision to hold a person to account for what they did." (Jefferson, 2019, p. 558) Similarly, many of the criticisms people make of instrumentalist accounts of blame are much like those made of consequentialist theories in general, and I will discuss some of these below. However, there are some criticisms of instrumentalist accounts that I believe do not apply to my theory, because they target instrumentalism as a theory of *blame* rather than as a theory of *blameworthiness*.

This set of criticisms argues that instrumentalist theories of blame produce an impoverished account of our blaming practices; that they fail to capture what we are really doing

when we blame others. This criticism comes especially from reactive attitude theories like Strawson's. To quote him "But this [the practical efficacy of the practices of blame] is not a sufficient basis, it is not even the right *sort* of basis, for these practices as we understand them' (Stawson, 1962, emphasis added). It is argued that instrumentalist accounts fail to capture something important about the phenomenology of blame (Wallace, 1994) (Shoemaker, 2015). Certainly sometimes we censure others in hopes of modifying their future behavior, but this is not always true of blame. And even in those instances where that aspect is present, there is much *more* to blame than that. There is resentment, a change in dispositions towards the person blamed, and many other aspects of blame as well. This is why critics of instrumentalism about blame tend to see such theories as having an impoverished view of blame. Blame isn't merely a means to some end; it is in itself an important part of our emotional lives and our interpersonal relationships.

That last is especially important because people (quite understandably) resist any theory that treats our interactions with other people as merely instrumental. Most theories of blame place great emphasis on the interpersonal nature of blame, with some, most famously Scanlon (2008) (2013), directly identifying blame as an impairment of a relationship (though it should be noted that Scanlon has a rather particular and fairly broad understanding of a 'relationship' as being a set of dispositions towards others). Others regard blame as a kind of moral protest that we make to others (Hieronymi, 2004) or argue more generally that our practices of holding others responsible are primarily a means of communicating something about the wrongdoing (and the wrongdoer) (McKenna, 2012) (2013) (Macnamara, 2011) (2015). And it is in the case of relationships these arguments against instrumentalism are at their most forceful. If my blaming someone is *just* an instrument for changing their future behavior, that seems disrespectful not

only of any victims but also of the agency of the wrongdoer, as we are treating them merely like a Skinner box to program and not a *person* who has done wrong.

But these are arguments against first order instrumentalism about blame, not second order consequentialism about blame. I am not here laying out an account of the phenomenology of blame, but rather an account of the justification underlying our theory of blameworthiness. That is, my theory does not describe *what* we are doing when we make moral ascriptions, but is an evaluation of the *criteria* we ought to be using when we are making said ascriptions. If you'll recall the arguments I made in Chapter 4, it isn't merely that SOC can evade this criticism, but that arguments like these are a large part of why I was motivated to move to *second* order consequentialism in the first place. For me, as for Hooker, the fact that meaningful relationships require us to *internalize* and not merely accept rules, and to bind ourselves to certain practices or obligations even if doing so leads to bad consequences in singular instances, is a large part of the reason we are second order and not first order consequentialists. For those who are drawn to consequentialism but do not like viewing our relationships with others (or our promises, or our special and role based obligations, etc.) in an instrumentalist and impoverished fashion, SOC allows us to get the best of both worlds. As a second order theory, my theory is not committed to a particular account of what blame consists of, but rather is compatible with most theories of blame, including Strawsonian theories. At least, this is true in principle; in practice, it would depend on what blaming practices would lead to the best consequences in the long run.

The constitutive features of blame are not what the theory seeks to explain, but rather part of the consequences to be considered when developing a standard of blameworthiness. Let me give an example to clarify what I mean: to go back to the discussion on Demandingness in the previous chapter, part of the second order reasoning for wanting our moral theory to not be too

demanding is that it would lead to there being more resentment in the world than we might want. One might reasonably be worried that a first order consequentialist account of blameworthiness cannot succeed because our ascriptions of blame are grounded in moral sentiments we cannot easily give up (McElwee, 2010) – the Strawsonian argument. But SOC takes into account our moral sentiments as part of its consequentialist calculus. It also assesses them as well of course, and may advise us to revise or even abandon some of them – but mindful of the realities of human psychology, it will advise gradual change rather than ignoring our sentiments entirely (*Ibid*.). This goes back to the idea of internalization we discussed in chapter 4: the first order theory recommended by SOC must be one we *can* internalize, and so must take our moral sentiments into account.

Now what is true is that instrumentalists about blame tend to blur the line between orders, and I think that they do so to their detriment. They would be better served, in my view, if they were to be clearer about the distinction I am here making, as it would defuse many of the strongest objections to instrumentalism. What is also true is that as I said back in Chapter 4 most of the benefits of second order consequentialism can be gotten with a sufficiently sophisticated indirect consequentialism, and it does seem like at least some instrumentalists are trying to be more indirect in this way, such as Jefferson, who takes inspiration from Railton's (1984) indirect (or as he calls it, self-effacing) consequentialism.

> Even if the ultimate goal of our responsibility practices is to change behaviour and moral sensibilities for the better, an indirect approach may be better suited to achieving this aim. We don't have to have the goal of improving behaviour in mind every time we make a judgement about blameworthiness or blame somebody. If we don't keep our eye on a consequentialist goal in every single instance, this may have better effects on the development of moral agency and it will also allow us to have more personally involved relationships, a worthwhile outcome in its own right. Whether self-effacing instrumentalism leads to better results is in the end an empirical question, but it is

> certainly more compatible with close interpersonal relationships than an explicitly instrumentalist, detached approach. (Jefferson, 2019, p. 561)

Vargas (2013) similarly also calls his theory 'second order' making moves that are similar to mine in order to evade the traditional objections to instrumentalism and emphasize the importance of internalizing norms which are generally efficacious rather than specifically considering whether some particular instance of praise or blame is instrumentally valuable (pp. 172-175). Unlike me, however, Vargas is not a second order consequentialist more generally, and this causes some problems for his theory that we will discuss later.

For reasons I outlined back in chapter 4, I prefer to draw a sharp distinction between orders rather than take the indirect approach, even though these two types of theories have similar practical implications. Either approach can also accommodate most theories of blame. It is even possible to recast Strawson's own theory as a kind of indirect consequentialism (McGeer, 2014), though it should be noted that it *is* a fairly substantial recasting. As I mentioned in previous chapters, it is my view that a consequentialist version of Strawson's theory lacks what I take to be the essential element of 'Strawsonian' theories, namely the emphasis on the *inescapability* of our emotional responses and reactive attitudes. Similarly, most functional theories of blame, which communicative theories are a subset of (Tognazzini & Coates, 2020), can be recast fairly easily as indirectly instrumentalist theories of blame. While doing so would, once again, be a recasting and not a mere restatement of the theory, it still shows how indirect instrumentalism about blame can accommodate other theories of blame.

The point is that second order consequentialist accounts of blame, or indirect instrumentalist accounts of blame, do not get the phenomenology of blame wrong despite what their critics might suggest. They are describing instead a different level of appraisal, engaging the question of blameworthiness and not the question of what the best version of our blaming

practices might be. In principle they are compatible with many theories of the latter. The question of what makes for the correct theory of our blaming practices is largely separate, and certainly separable, from my purposes in this chapter, and so I will for the most part set it aside. It is at least in part an empirical question, and more a matter for philosophy of mind (and psychology) than ethics.

That said, it is worth noting that many prominent theories of blame, including Strawson's, Scanlon's, and many others, are sometimes ambiguous about whether they are giving a descriptive or a normative account of blame. The dialectic of blame is often confused. For instance, some discuss blame in the context of free will, and how we can justifiably blame others if determinism is true (Stawson, 1962), while others seem to be laying out the proper role of blame in our interpersonal relationships (Scanlon, 2013) (Wolf, 2011), yet others try to find what it is we are renouncing when we forgive someone (Hieronymi P. , 2001). In other words, blame is often discussed in the context of a larger agenda, and this chapter is no exception. My account is not at all descriptive, so the fact that we need not have any second order consequentialist reasoning behind our *actual* blaming practices is not a problem for it. That said, since instrumentalist accounts of blame do straddle the line as I described, while this particular set of objections to instrumentalism I believe has little force against my theory, there are other objections to instrumentalism that do apply to my theory. These will be the subject of the next few sections.

The conceptual objection

The second sort of objection that arises with both instrumentalist and consequentialist theories, regardless of directness or indirectness, is the same sort of objection that always arises with such theories, and has to do with their nonintuitive conclusions. This objection can be

further divided into two parts, which I call the conceptual objection and the counterfactual objection. The first is the worry that consequentialist theories represent some features of blame as being external to blame itself and based on contingent factors, when they are in fact inherent in our concept of blame. The second is the worry that consequentialist theories of blame can allow for counterfactual situations where we are justified in blaming the clearly innocent, or not justified in blaming the clearly guilty, or otherwise have nonintuitive results. It is worth noting that these objections are clearly related, as a large part of the reason for being worried about allowing blame and desert to come apart *conceptually* is that it leaves open the possibility of situations where they can come apart in practice. Still, I will treat them separately.

We discussed the conceptual objection last chapter when discussing Norcross, who argued against defining "wrong" based on the consequences of doing so, because our concept of wrong is constrained by certain principles which that definition does not adhere to. McElwee extends this argument to blameworthiness, pointing out that the principles Norcross highlights constrain both concepts and not just one. In addition to those two principles – universalizability and allowing for imperfection – we might also add other principles. For example, we might hold that if someone has not *done* an action they should not be held responsible for that action. We may also want to impose some kind of proportionality constraint. Regardless, the objection is that the definition of blameworthiness I offer here does not conform to the principles our concept of blameworthiness is in fact constrained by. Hence why I call this the 'conceptual' objection.

The debate here has some similarities to the externalist-internalist debate in metaethics. A judgment internalist about moral reasoning holds that someone cannot actually make a moral judgement without being in some way motivated to act on that judgement. An externalist, by contrast, holds that the relationship between making a moral judgement and being motivated to

act on that judgement is merely contingent (Rosati, 2020). In both cases, the conceptual and counterfactual objections overlap, as the easiest way to understand the difference is by asking whether the connection in question is contingent or necessary. But I think there is a conceptual objection that is slightly different than the worry that a theory may allow for nonintuitive cases, as there is also a worry, that exists at the level of theory rather than practical implications, about getting the concept itself wrong. A judgement internalist undoubtedly considers it a troubling fact for externalists that their view allows for someone to exist who can make sincere moral judgements without being motivated to act on them. But they would also argue that such a person would be not 'getting it', that they are not actually making sincere moral judgements in the first place. Part of the concept of being a good person, one might claim, is that you are motivated *non-derivatively* by justice, equality, and so on, rather than motivated by doing the right thing but where the content of 'the right thing' is filled in separately (Smith, 1994, pp. 72-75).

Similarly, someone might critique a second order consequentialist theory of blameworthiness as not actually getting the concept of blameworthiness right if it makes it out to be separable, even just in principle, from proportionality, agency, and so on. Someone who holds another person blameworthy without any regard to being proportionate to their wrongdoing isn't just making a mistake in judgement but cannot be correctly said to be holding the other blameworthy, merely (perhaps) unjustly blaming them. I will say that this argument seems *prima facie* somewhat weaker to me in the case of blame, though we must be careful to distinguish blaming someone (which can be irrational and unjustified) from correctly judging them to be blameworthy. The conceptual objection here is that my theory seems to allow that in some

circumstances we can *correctly* hold others to be blameworthy out of proportion to their wrongdoing. And that, the critic might say, is to get the concept entirely wrong.

However this still seems weaker to me than the judgement internalist argument. Someone who makes judgements of blameworthiness without reference to proportionality may be doing a bad job of it, but they still seem to be *making* such judgements, rather than doing something else. By contrast, there does seem to me an important sense in which someone who claims to think that murder is wrong but is in no way motivated to not commit murder doesn't *really* think that murder is wrong. In addition, I think the historical contingency is weaker in the case of blameworthiness. There have been many times and places where people have not adhered to principles like proportionality in their blaming practices and while we may think that they were *wrong* we generally do not think that they were engaged in a different practice altogether.

Still, I think the critic of consequentialism can press the point, and argue that when we judge people blameworthy of some wrongdoing, considerations of proportionality and agency *must* be tied into it or we aren't really judging them blameworthy. Perhaps the easiest way of bringing out this point is to think of the case of dangerous animals. When we judge some man-eating tiger to be a threat and resolve to trap it or kill it, what we are doing is similar in many ways to what we do when we judge someone guilty of a wrongdoing, but because animals do not have the right sort of agency what we are doing seems to be something distinct from judging the tiger to be *blameworthy*. This suggests that recognizing the wrongdoer as an agent, at least, is intrinsically tied to judging them blameworthy.

Now, of course, I do believe that the extent to which we blame someone should fit the crime. I do believe that intent matters, that someone without control over themselves should not be blamed for their actions, and so on. I believe in the constraints of universalizability and

allowing for imperfection. I have argued some of these very things in the last few chapters, and will go on to argue for them in the next few. But I believe these things because I believe that the best consequences in the long run are obtained when the theory of blameworthiness we internalize takes intent into account, does not blame people who are not properly responsive to reasons, allocates blame proportionally to the crime, and so on. And I believe that internalizing a theory of blameworthiness that does *not* do these things will lead to worse consequences in the long run compared to internalizing a theory that does. However, I do not think these constraints are inherent in the concept of blameworthiness – rather, they are the result of my theory as applied to the real world. Depending on how one looks at it, this can be seen as both a positive point and a negative point.

The negative point is this: if you think that intent mattering or proportionality or so on are inherent features of blameworthiness, then my theory is something you are going to regard as incomplete or incorrect. If you are worried about a theory that says that blameworthiness *can* be disproportionate to the crime, that remains to some degree a worry even if the theory doesn't *actually* say that it ought to be. The positive point, though, is that it gives a reason and *justification* for proportionality, for the importance of agency and so on, that doesn't simply amount to 'because that is what being blameworthy means'. In fact, I would argue that my theory ties blameworthiness with agency very tightly indeed, but that it does so because of how the theory explains and partially defines agency, rather than because it builds the connection between the two into the theory's premises. The price of having a justificatory story for the connection between blameworthiness and intent, agency, or proportionality, is that the theory cannot make these connections inherent to the concept itself.

Further, because it is a *consequentialist* theory, it also does not make the connections necessary, but rather contingent on facts about the world, facts about what practices of blame-holding actually lead to the best consequences. Some other nonconsequentialist second order theory could still give an explanation and a justificatory story about the connection between blameworthiness and agency, but might instead make the argument that the connection was a necessary one. If you are worried about the two coming apart even conceptually, then that other hypothetical theory might well be more preferable. I am obviously not here trying to argue against my theory in favor of that hypothetical other theory – indeed, I do not know what such a theory would look like – but rather am acknowledging that my theory does lose some plausibility points here. It makes what many might think are conceptual truths into outcomes of the theory that are contingent on empirical facts about human behavior and responses.

It is possible to modify my theory to account for this in some ways, just as many consequentialist theories can be modified to take into account distribution, desert, or similar factors. The solution is, of course, to build those factors into one's theory of the good. Take as an example the debate between total utilitarians and those that disagree with them. Total utilitarians argue that the best world is the one with the greatest total welfare, regardless of how that welfare is distributed. The reason, they claim, that it seems like we care about the equitable distribution of welfare is because we confuse it with things like money, resources, or leisure: the things that in most real world problems of distributive justice are the things being *distributed*. These things have a diminishing effect on welfare, so the greatest total welfare is produced when these things are distributed equitably. Thus, the total utilitarian still believes that the best worlds are those with equitable distributions of resources, arguing that the reason for that is precisely that that is what leads to the greatest total welfare.

But other utilitarians – and other types of consequentialists – are not satisfied by that. They think, upon reflection, that even welfare must be distributed equitably. They dislike that the total utilitarian's argument for equity relies on contingent factors about the world (namely that most resources have diminishing returns in their effects on welfare). Or they worry about hypothetical 'utility monsters' (Nozick R. , 1974, p. 41): individuals that have no limit or diminishing return to the amount of welfare they gain from resources. The total utilitarian seems committed to saying that were such a being to exist, the right thing to do would be to give them all the available resources, as that would maximize the amount of happiness in the world. Regardless of the realism of such a scenario, many people are moved enough by the claim that such a thing is unacceptable in principle that they reject the total view, adding some notion of fairness or desert into their theory of the good (Broome, 1991) (Feldman, 1997). The case of my theory and blameworthiness is very similar. Indeed, one can find many similar cases throughout the consequentialist literature, as the debate over what belongs in the theory of the good and what is an instrumental good derived from the consequentialist theory is a perennial one.

Thus far, I have been mostly silent about what theory of the good I prefer. This is because my main goal was to argue for the benefits of SOC independent of the theory of good that accompanies it. There are many plausible versions of second order consequentialism, as I see it, that can be paired with different theories of the good. For the rest of the chapter I will provide a defense of those versions that do not incorporate desert or similar concerns into their theory of the good, but that should not be taken as a sign that those are the only versions I consider plausible or worth expanding on.

To some extent I *do* prefer them: as I stated in earlier chapters, one of the main reasons I am drawn to consequentialism is that it provides a relative simplicity and unity of explanation.

Thus I prefer it if more of our moral truths can be derived from the theory, rather than incorporated into the premises of the theory. A consequentialism that derives desert, equity and so on from a total, maximizing consequentialism of a unitary or at least a simplified list of goods has much more of the sorts of theoretical virtues that draw me (and, I suspect, many others) to consequentialism. Despite that, I do not rule out the possibility that the correct theory of the good might need to be quite complex, as I think there are still many reasons to believe that that may be the case.

But for the sake of maintaining my overall claim that second order consequentialism provides a fruitful line of pursuit independent of what theory of the good you use, I will in this chapter try to defend that variation that is the hardest to defend: the one where desert, equity, agency, and many other things are *not* part of the theory of the good. If a defense can be made for this variation, all other variations of SOC will also benefit.

As this is simply one instance of a perennial debate among consequentialists, it should be no surprise that the outline of the argument for both sides is well established (Kagan S. , 1998, pp. 48-59). To a large extent, how worried you are about the conceptual objection hinges on how worried you are about counterfactuals. It is possible to worry about the concept of blameworthiness being separated from things like proportionality for mainly theoretical reasons – that is, 'theoretical' as in pertaining to the nature of the theory. Concerns like simplicity, elegance, unity of explanation, and other theoretical virtues or vices might pull you towards one camp or the other. However, most of the people who worry about instrumentalist or consequentialist accounts of blame and blameworthiness worry about the *conceptual* gaps because they fear it could lead to *justificatory* gaps. That is, cases where the theory might say we are justified in blaming someone who has done no wrong, or perhaps someone who lacks

agency. Just as with those who reject total utilitarianism due to worries about utility monsters, many of these objections arise from the possibility of troubling counterfactuals, which brings us to the next part of the objection.

The objection from counterfactuals

Possibly the greatest worry about my theory of blameworthiness, and perhaps any theory of blameworthiness that incorporates consequentialist reasoning, is that one can imagine situations in which the theory justifies moral judgements that are not merely nonintuitive, but outright unacceptable. For instance, it could be argued that my theory would imply that the punishment should be grossly disproportionate to the wrongdoing, if it were the case that that would create a deterrence effect that would lead to less wrongdoing in the future. Similarly, and perhaps much worse, one could imagine a scenario where it would result in better long-term consequences if we adopted as a practice blaming the *victim* of the wrongdoing instead of the perpetrator. Perhaps, for instance, doing so would encourage people to take much stronger measures to protect themselves, resulting in a decrease in wrongdoing overall.

Different sorts of theories will lead to different potential problems, and what may be an issue for one consequentialist theory of blameworthiness might not be for another. For instance, McGeer (2015) worries that because Vargas's (2013) account has a different theory for what justifies our ascriptions of responsibility and what makes a moral agent, there is a danger of a 'justificatory gap'. Vargas is, like me, a second order rather than a first order instrumentalist about blame. He, also like me, thinks that the best account of agency is some type of reasons responsiveness: that is, he think that for something to be able to held morally responsible it must be responsive to moral reasons. But Vargas explicitly declines to embrace broader consequentialism, wanting to remain neutral about substantive normative commitments outside

the area of responsibility (*Ibid.* p. 128). But this means that that his account of agency is *independent* of his instrumentalist theory of blame and this, McGeer charges, leads to the possibility of cases where we are justified in blaming individuals according to his instrumentalist theory, who are according to the independent theory of agency not agents. Jefferson (2019) worries that her own account might have the same problem.

By contrast, this is not a worry that arises on my account. This is because Jefferson and Vargas, though instrumentalists about blame, are not consequentialists in other areas (*Ibid.* p. 570). I, however, am *always* a (second order) consequentialist. Thus, my theory of agency *also* has a consequentialist justification, so agency and being susceptible to ascriptions of responsibility cannot come apart on my view. I believe that an actor has agency precisely if it is the case that holding it morally responsible leads to the best consequences in the long run. Thus, unlike Jefferson and Vargas, I do not worry about any justificatory gap because I make the standard move classical instrumentalists do and define agency in terms of moral responsibility. There will not arise a McGeer type scenario where responsibility norms justified by second order instrumentalism might be incorrectly applied to non-agents in some individual cases, because in my theory those very same norms also underlie agency.

This seems to imply that if I believed that holding babies, animals or machines morally accountable – i.e. blame or praise worthy – led to the best consequences, I would also believe they had agency. This is an implication that I do not deny. However, recall our discussion from chapter 5 (pp. 73-75): for SOC to actually recommend holding (say) animals morally accountable it would have to be the case that doing so led to the best consequences not merely in some limited circumstances but in the vast majority of circumstances, and that it led to the *best* consequences not merely some good ones. And specifically, that it led to better consequences

than treating animals the way we treat them now (as not morally accountable, for the most part). It is my firm belief – supported by my best understanding of agency[32], the natural world, and how minds work – that the only way for it to be *true* that holding an animal morally accountable would lead to the best circumstances in sufficiently general circumstances would be if that animal were responsive to moral reasons in the right sort of way. I am comfortable saying such an animal has agency: I believe that at the present moment no nonhuman animal satisfies that criteria.

Instead if the problem that agency and blameworthiness can come apart, the worries about my account are standard worries about consequentialist accounts of responsibility, that they might justify blaming the innocent or even the victim or blame out of proportion to the wrongdoing. The discussion around these sorts of cases bears a great deal of similarity with classic problem cases for consequentialism such as Transplant, which we have already discussed (see above). It is possible to engineer scenarios where consequentialist theories can have outrageously unacceptable recommendations, but a classic response is that this is only a true problem for consequentialism if those scenarios are actually realistic or plausible. After all, the more implausible or unrealistic the scenario, the more reasonable it is to assert that we cannot trust our judgements about it. We have good reason to think that our intuitions and judgements are at least less reliable than normal when it comes to highly unrealistic scenarios, because they are calibrated for realistic scenarios (Kagan S. , 1998, pp. 76-77).

Furthermore, the scenarios where SOC leads to unacceptable results are more unrealistic than Transplant or similar cases. First order consequentialism can lead to unacceptable results in

[32] If you were to ask me what agency *consists* of I would likely answer with some version of Moderate Reasons-Responsiveness, as outlined by e.g. Fischer and Ravizza (Fischer & Ravizza, 1998). If you were to ask me what *makes* an agent morally responsible, then I would answer with SOC.

individual cases, if the circumstances are sufficiently specified to remove all possible long term or unintended consequences or any other consideration an indirect consequentialist would normally point to as support for the claim that their theory would not lead to unacceptable results in realistic cases. But mine is a second order consequentialism, like rule consequentialism is. I think we should adopt that moral system for which it is true that the best results would come from adopting it in the long term – and to adopt a moral system is to internalize it, to bind and commit oneself to it. In individual cases, that moral system is not necessarily always going to recommend the action with the best consequences. And further, internalizing a theory of blameworthiness has consequences over and above those of simply blaming as that system of blameworthiness says we should, as we discussed in Chapter 5 (p. 75).

Rather than using first order consequentialism on individual instances of blame to see whether or not they are merited, second order consequentialism is about using consequentialism to evaluate, interrogate, and if necessary revise our blaming practices as a whole. Thus, for it to be the case that my moral system recommends unacceptable acts, it must be *generally* true that those acts lead to the best consequences, not merely true in some individual instances. And it must be true that adopting and internalizing the principles that allow for such acts will lead to the best consequences. For *that* to be the case, it is not sufficient to simply set up some highly specified scenario involving individuals; rather, one has to imagine a different world entirely, with either an entire society that is both very strange and itself invulnerable to change, or where some very basic features of human psychology are different.

To look at the case of proportionality and desert, it is not that hard to imagine a situation where *some* good results might plausibly come from assigning disproportionate levels of blame for wrongdoings in *some* cases. But my theory is maximizing on the second order level: it is not

sufficient to cause a problem for my theory if there is a concept of blameworthiness that is morally troubling but leads to good consequences. For it to be a problem, it must be the case that it would lead to the *best* results for us to hold others and ourselves blameworthy disproportionate to our actions. And not merely that it would do so in some isolated instances, but that it would do so in the long run, meaning at least in the majority of cases.

And since adopting a disproportionate principle of blame seems likely to itself have adverse consequences, this makes it all the more unlikely that such a principle could lead to the best consequences overall compared to adopting a proportionate principle. Furthermore, principles like proportionality aren't to be considered in isolation but as part of a complete theory of blameworthiness. Even if one can imagine circumstances where disproportionate blame has good consequences, it seems implausible that the *best* theory of blameworthiness overall will not include a proportionality constraint. Even if we granted that disproportionate blame had good consequences because it deters more effectively – and as I'll note below, I think we have very good reasons to *not* grant this – including it into our overall theory of blameworthiness will undermine *other* aspects of it. For example, it will make attempts at restorative and reformative justice have worse consequences. When all this is taken into account, all plausible candidates for the overall best theory of blameworthiness will include a proportionality constraint. As such, so I do not regard disproportionate blame as a serious problem case for my theory.

It is worth going into detail on why I find this so implausible. Interestingly in part it is because I *don't* think that people find it immediately apparent that holding people blameworthy disproportionately would be wrong. Throughout history and even the present in many places, people have tried disproportionate punishments for crimes, or treated what I regard as minor vices with an excessive degree of social opprobrium, or engaged and engage in victim-blaming. I

believe that these practices are wrong, but I did not arrive at this belief from nothing. Rather, I believe this in no small part because I see the effects these practices have had in real life. As I discussed in earlier chapters, I think our intuitive judgements are in fact formed – and informed – by our knowledge of history and the impact of certain moral beliefs in the real world. I am in fact very much of the opinion that it is wrong to hold people to blame out of proportion to their wrongdoing, but this is in part *because* I am very confident that any plausible SOC would show that there doing so would not lead to the best consequences. My confidence in the former claim and my confidence in the latter claim are the same, because in both cases it ultimately comes from the same place: the combination of my moral sense, my understanding of humanity, and my knowledge of the world. Conversely, my confidence in one cannot be undermined without also undermining my confidence in the other. This is why I am more willing to bite the bullet in cases where I think SOC *would* have us revise our longstanding opinions about blameworthiness, as I talked about last chapter, because these are also the cases where my confidence in those longstanding opinions is much less. Thus, I do not believe that there are realistic scenarios that are genuinely troubling for my theory, such that for example it would say that we should regard others as blameworthy out of proportion to their wrongdoing.

The situation is even more stark with the hypothetical scenario where blaming the victim leads to the best consequences. It seems deeply unlikely to me that there are remotely plausible cases where we are justified in blaming the victim because the *practice* of doing so would lead to the best long term results. It is of course easy to imagine scenarios were blaming the victim leads to the best results because of the peculiar circumstances of some individual case, but that is not sufficient for SOC to recommend adopting it as a general practice. It is also possible to imagine a way in which adopting victim-blaming as a general practice might lead to *some* good

consequences (e.g. by encouraging people to act in ways so as to avoid becoming victims) but that is also not sufficient for SOC to recommend it. For SOC to tell us to adopt the practice of victim-blaming, doing so must lead to the *best* consequences overall when considering *all cases*, and must be superior when compared to every alternative principle that might also have good consequences,

Once one understands this, second order consequentialist reasons to not adopt the practice of victim-blaming abound. For one thing, is victim-blaming the *best* way of encouraging people to avoid becoming victims? For another, is encouraging people to avoid becoming victims the best way to deal with the underlying problem as opposed to, e.g., discouraging people from turning others into victims? Of course we can try to do both, but what if the practice of victim blaming undermines our effort to do the latter? And that is not even taking into account that internalizing a principle of victim blaming has negative consequences in and of itself at both the societal level and the individual level – if nothing else, it has a direct adverse effect on the mental health of victims.

And again, these aren't wholly hypothetical scenarios – over the course of human history many societies and cultures both in the past and the present have adopted some victim-blaming principles in some form and my judgment of the consequences of doing so is based on these examples. We know that adopting a principle of victim blaming has adverse consequences at the societal level because we've seen it. We know that *internalizing* a principle of victim blaming has adverse consequences at the individual level because we've seen it. I have reasons for not believing that victim blaming is morally justified, and many of these reasons are the *same reasons* I believe SOC will not recommend victim-blaming in any remotely realistic circumstance. Taking into account the consequences of adopting and internalizing a principle of

victim blaming, and comparing it with other principles that may accomplish similarly good consequences with fewer adverse downstream consequences, I do not have any serious worries that my theory will recommend blaming the victim in any but the most distant possible worlds.

All even slightly plausible scenarios where the best results come from internalizing the principle of blaming the victim for a wrongdoing involve highly artificial circumstances, usually involving mind control or carefully designed societies (and, of course, one might argue that such circumstances are by definition *not* 'plausible'). Even in those cases, though, it still seems to me that the best results would not come from blaming the victim in the long run, but rather in blaming the mind controller or the society. Let us imagine a society that is, by whatever means, set up so that blaming the victim will lead to the best results in the majority of cases and so in the long term. The unsaid but still very important qualifier here, though, is that is true for so long as that society continues to exist and to be so set up. That is, it is true only in that society. But SOC does not relativize by society, but asks what first order moral theory we should adopt on the basis of which one will lead to the best consequences in the long term. Surely any such theory at all worth considering would not be one that sets aside the possibility of reforming and changing the societies we live in – that is, the question of what the society we live in ought to be like is also within the scope of the moral theory. The question of collective responsibility will be the subject of a future chapter, but for here it is only important to note that we surely have some sort of responsibility to better the society we live in. Even if it were true that within the society as it exists it would lead to better results to blame the victim, for such a case to be a problem for SOC it would also have to be true that my theory would justify setting up and maintaining such a society in the first place, and not overthrowing or resisting it. But SOC would only do so if the best consequences in the long run were achieved in a society that rewards blaming the victim,

and that seems unlikely to me in the same way that it is unlikely at the individual level; if anything, it seems even more implausible.

To truly tighten up the example, we have to imagine more than just a strangely organized society, as we have to fix things *outside* the scope of the consequentialist theory. It must simply be the case that the world is such that blaming the victim leads to the best consequences and there is *no way to change that*, or at least no way to change that that would not itself have irredeemably bad consequences. In other words, we take the Pettit and Smith approach (2000) and imagine a world in the grip of some mad scientist who insures that our theory has nonintuitive conclusions. Indeed, I would go even farther and imagine some omnipotent but evil demon that, for whatever perverse reason, has set up the world such that blaming the victim is what leads to the best consequences in all (or at least the vast majority of) cases. Such a scenario is very extreme, but I also think it is the most problematic for my theory because I see no option but to bite the bullet here. Given such a scenario, SOC *would* recommend internalizing the principle of blaming the victim. In general, SOC would recommend bowing to the whims of an omnipotent deity, and would also seem, somewhat more uncomfortably, to render any such being immune to moral judgement (because no judgement of them could possibly have consequences of any sort; one might as well morally judge natural laws). Fantastical though this scenario is, this is a very troubling result, since surely it is the case that even if we cannot resist such beings, we are at least able to morally judge them. But, at least for people living in that world, they cannot.

This is a genuinely troubling conclusion but I do have some arguments to at least blunt its impact – to limit the loss of plausibility points, as Enoch would say. The first is to argue that *of course* my reaction to the presented scenario is to continue to hold that blaming the victim is

wrong and that the all-powerful being described is evil. That's because the moral system I have adopted is the one that SOC tells me to adopt, the one that has the best results in the world I actually live in. The fact that in this alternate world the moral system with the best consequences would involve blaming the victim does not alter that I must judge the world by the moral system that I have adopted.

This line of argument, however, is one I dislike. I am not fond of this kind of rigidifying trick when it is done by perspectivalists in metaethics, and it would be rather hypocritical of me to embrace it here. It is true, in any event, that while this argument can reconcile the fact that I continue to hold that it is wrong to blame the victim while also allowing me to admit that in such an alternate world SOC would tell me to do otherwise, it does not address the core of the issue. Namely, that we have a conviction that blaming the victim is never the right thing to do regardless of contingent facts. Rigidifying does not allow us to actually blunt that conviction, merely to do a sort of end run around it.

The second line of argument, then, is to actually try to face that conviction head on. To ask us, in effect, to understand that the people of that world that blame victims of wrongdoing *are* doing the right thing, given the circumstances that they are under. The argument here is that while a moral theory might say something about ideal circumstances, when it comes to telling us what to do in any actual situation it is bound by the constraints we are operating under. To give a concrete example, if I am faced with two ill patients but have medicine enough to save only one of them, I am not wrong for letting the other die. This is true even though it is also true that letting someone die when you have the means to save them is obviously wrong. Something that would be wrong when considered outside of context becomes acceptable when it is understood that one is operating under the constraints of limited resources.

The people living in the hypothetical world described above are similarly acting under the constraints of their circumstances when they blame the victims of wrongdoing. This is not meant to say that collaborating with an oppressive system is acceptable or moral (though I would want to say that doing so is at least *sometimes* forgivable). The people in question are not living in a society that can be resisted, defied, fought, or reformed. By stipulation, the fact that blaming the victim leads to the best results is impossible to change, more akin to, as I alluded to earlier, a species of natural law. In that situation, it seems to me that were I placed in it I would argue that blaming the victim is the right thing to do, but only because of the constraints of the world as it was. Just as allowing people to starve is morally acceptable only if there is genuinely not enough food to go around. When I see things from that angle, I can think that the people in that world that blame the victims of wrongdoing aren't wrong, but rather are making the best they can of a bad situation. At least for me, this allows me to be reconciled to the fact that my theory creates an absurd conclusion in that admittedly very extreme and unrealistic scenario. And it should be noted that the scenario in question is so extreme that *everyone* will need to bite some bullet. To hold that it is wrong to adopt the principle of victim-blaming in a world that is set up like this one is, I think, like holding that it is wrong to steal food even to feed starving children. We all do care about consequences to some extent, and scenarios where the consequences are *so* divorced from our ordinary moral intuitions as this are genuine moral dilemmas where any choice at all involves swallowing some unpleasant implication. There being no option here that is all around congenial, I choose to stick with my consequentialist commitments.

For it to be plausible that the best results come from blaming the victim in the *absence* of outside control and design, the agents involved must respond to moral reasons in so different a fashion to anything we are familiar with that they are effectively alien to us. When faced with

such beings, I find myself unable to put any faith in my pre-theoretic intuitions about blame and responsibility, even such a basic one as it never being acceptable to blame the victim. As I've discussed before, our notions of blame and responsibility are already relativized by species. When we are dealing with a group of agents who are so different from us, it would be prudent to treat them as though our notions of blame and responsibility do not apply, not for the same reason we do so for animals (these hypothetical agents are responsive to moral reasons, unlike nonhuman animals) but in the same way that we are adjusting our concept of blameworthiness to take into account their different responses.

All this does not discount the possibility of much more plausible scenarios where second order consequentialist justifications of moral judgements lead to nonintuitive results. I have already mentioned two areas where my theory has led me to believe that we need to alter the way we evaluate certain cases: firstly, that we have historically been too harsh on cases of addiction; secondly, that we are too undemanding on those with wealth to give away. One might think that not *too* much would need to be different about the way the world worked, the way society worked, or the way people acted for my theory to recommend the opposite. But I am not troubled by this possibility: I am willing to bite the bullet here and say that in such a world my recommendations for how we should alter our evaluative criteria ought to, and would be, different. This is not something I see as a large problem with my theory, whereas it very much would be a large problem if there are remotely plausible (as opposed to merely possible) scenarios where my theory would recommend that we blame the victim of a wrongdoing.

Perhaps the best way to sum it up is to put it in terms of possible worlds. I would consider it a grave problem for my theory if there were close-by possible worlds where my theory led to absurd results like that the right thing to do is to blame the victim. I consider it much less of a

problem for my theory if it leads to absurd results only in very distant possible worlds. It is important to note that my theory is a second order theory, that asks us to adopt that first order moral theory that would have the best consequences and to judge individual instances within the framework of that moral theory. When I take this into account, I find it highly implausible that there are any nearby possible worlds where my theory leads to results I find unacceptable, because even if one can concoct some artificial scenario where it might individually lead to the best results in that scenario to do something absurd it would not make it so that that absurd action would lead to the best consequences in enough cases that my theory would actually endorse it.

Conclusion

My theory of blameworthiness is unquestionably revisionary. It does not intrinsically tie blameworthiness to intentionality, will, or motive (though it does *instrumentally* do so) which is quite radical. However, if that theory of blameworthiness allows us to answer otherwise unanswerable questions about how demanding our moral theories need to be, I think it is deserving of serious consideration. And when considered as a *second order* theory, it is much less revisionary than one would initially think when faced with the proposal that someone is blameworthy only if it is the case that it would lead to the best consequences to blame them. This is because the theory is then also taking into account the consequences of internalizing that judgement of blameworthiness, which are over and above the consequences of the judgement itself.

The theory makes what many would consider essential features of blame into derived features based on contingent facts. But that is true primarily at a second order level. When considered at the first order level – that is, when we consider the actual practice of blaming and praising that the theory would recommend – we see that many of those features emerge and are

still as tightly tied to blame as we would expect. Since my second order theory is also used to define agency, for instance, my theory makes agency and blame inseparable, which is of course exactly what we would expect. While there still exists a sense in which those features ‘could’ come apart, that possibility is so remote that I do not find it worrying. My theory is revisionary, and I have come to change and adjust my own standards of blaming in the light of it, but it has not asked me to give up any of the features which I think are truly essential to blame.

Chapter 10: Second Order Consequentialism and Collective Responsibility

Introduction

In an earlier chapter I argued that second order consequentialism would favor a less than maximally demanding system, because a moral system that is unreasonably demanding has too many bad consequences. It does not allow for people to pursue personal goals, it is impractical, and so on. In that chapter, I developed the argument that a consequentialist may be driven, by purely consequentialist (albeit second-order) considerations, to adopt a less than maximally demanding system. But any system that is less than maximally demanding runs into some serious problems when we expand beyond the level of the individual and start to consider the actions of groups and larger entities, leading to collective action problems. Actions that in that earlier chapter I described as 'suboptimal but not blameworthy' may lead to unacceptably bad consequences when taken by the majority of a large group of people. To put it another way, the collective result of a large number of people individually acting in non-blameworthy ways can result in a level of harm that would clearly far cross whatever threshold of blameworthiness we might have were it done by an individual. The question of how we deal with such cases, and the complications and problems that arise in such cases, are the subject of this chapter.

Polluter's Dilemmas and demandingness

The reader may at this point be somewhat surprised that this chapter exists – did I not, in chapter 5, claim that I rejected Collective Consequentialism in favor of the individual form? And then in chapter 7 I rejected CC as being a viable solution to the Demandingness Problem. As a result, it might seem that I would also reject notions of collective actions or collective responsibility. However, what I rejected in previous chapters was not the general idea of

'collective responsibility' but rather the specific way that Collective Consequentialism goes about incorporating it. CC takes a collective view of the *principles* that SOC (or rule consequentialism) tells us to adopt, and this form of collectivization I think leads to enough problems and has few enough advantages that I don't wish to adopt it. But that doesn't mean that I don't want to adopt some notion of collective responsibility into my theory. This chapter will explain how I wish to do that.

Let me start by motivating why I think my theory needs to incorporate some notion of collective action despite rejecting Collective Consequentialism. In fact it is partly *because* of my rejection of CC, as that rejection creates a problem that I believe I need a notion of collective responsibility to solve. I have developed a theory of blameworthiness that allows for blameless actions that are less than maximally demanding *without* (as CC does) distributing the responsibility for doing good equally onto each of us. That is, as I mentioned in chapter 7, I do not think – as for example Hooker might argue – that we each have the obligation do give X% of our income to charity and anything over that is supererogatory, where X% is determined by how much would be sufficient if *all* of us where to give that much. I rejected this kind of collectivization in chapter 5. But at the same time my theory is not maximally demanding. This creates problems when we discuss cases where many individuals, each of whom are acting in a non-blameworthy fashion, create outcomes which are clearly morally unacceptable. It is these sorts of cases I must now try and find a solution to.

Derek Parfit calls cases like these "Many-Person Dilemmas" (Parfit, 1987, pp. 58-62) but that is largely because he develops the concept from game theory. The way game theory (and thus Parfit) approaches this is cases where actions that are individually the 'best' are collectively self-defeating. Parfit takes this approach because at this point in his book he is interrogating our

reasons for acting. However, the problem I am interested in, while structurally similar to the cases Parfit or game theorists talk about, has to do with permissibility rather than our reasons for action. The issue is what theories consider morally acceptable on the personal scale leading to unacceptable collective results rather than individually optimal actions leading to collective suboptimality. These problems are also called "many-hands problems' (Thompson D. F., 1980) or simply *collective action problems* (Kagan S. , 2011). I will sometimes refer to them as the latter, but for the most part I will take a cue from a paper by Justin Moss (2011) and call the types of problems I am interested in *Polluter's Dilemmas*.

The reason for this name is of course because some of the most pressing cases of this type of problem come up with pollution. Let us take the simple example of food waste: almost all of us waste some amount of food on a daily basis. For the vast majority of us, this wastage is too small for it to be reasonable to hold us to account for it, and it does not rise to the level of blameworthy behavior. But on the large scale, this sort of behavior leads to massive food wastages in a world where people are starving. If our moral theory says that the behavior of each individual contributing to the state of affairs is morally acceptable, it would seem to follow that the state of affairs is morally acceptable. But, of course, the existence of massive amounts of food waste in a world that has starving people in it hardly seems like a morally acceptable state of affairs!

The other reason to call them Polluter's Dilemmas is to highlight, as this example shows, that these are not merely a theoretical worry, but a set of very real and very pressing problems that our moral theories have to address. Very similar analyses can be made for many other systemic and structural ills, such as patriarchy, racism, or economic inequality, but perhaps the most pressing problem of collective action facing us today is climate change. It is hard to argue

with the statement that a moral theory that does not somehow produce a duty or moral obligation to take action in the face of climate change and other global catastrophes is simply inadequate. Any remotely acceptable moral theory must hold that the states of affairs that arise from Polluter's Dilemmas are not morally acceptable, and hold further that we have a moral obligation to address and fix them. Yet at least *prima facie* it seems that any attempt to turn that obligation into prescriptions for action will result in unreasonable moral demands on the individual level.

Another way of looking at it is that we are speared on the horns of a larger dilemma. On the one hand we have the Demandingness Objection, which we discussed at length in earlier chapters, and on the other we have Polluter's Dilemmas. It is important to emphasize, however, that it is not simply a matter of finding the right balance, the sweet spot where our moral theory is demanding enough to prevent Polluter's Dilemmas from arising but not so demanding that it fails the Demandingness Objection. If that were all, there would be no problem, or at least the problem would be a lot more solvable. Rather, the issue is that there is a *mismatch*: any moral theory that allows for people to pursue their own projects, to be less than maximally perfect – that is, in short, not unreasonably demanding – seems to allow for the existence of states of affairs that are clearly morally unacceptable. And the reverse: a moral theory whose prescriptions preempt situations like food waste or climate change from ever arising must be one that is unreasonably demanding on the personal scale.

To be clear, in most real-life cases of pollution, oppression, or other systematic evils, there are absolutely people involved in the creation of these states of affairs whose conduct *unquestionably* rises to the level of blameworthiness. It is certainly not the case that the conduct of everyone that contributes to climate change is morally acceptable. What is the case though, and what elevates this to more than a mere theoretical worry, is that it is not *solely* morally

unacceptable conduct that results in morally unacceptable outcomes. If we were to hold to account everyone whose individual actions contributing to climate change rise to the level of being unquestionably blameworthy, and even if we took some sort of corrective action to address that, it would still not solve the problem. It would certainly *better* the situation, and ought to be done for that and many other reasons! But even if we removed all the cases of individuals acting in a blameworthy fashion we would still be left with a situation it seems clear that we have a moral obligation to do something about. And yet, it seems, it would be a situation where *no one in particular* has a moral obligation to act anymore, since it would now be the case that actions to alleviate the problem are merely supererogatory.

At this point it may be tempting indeed to go back to the maximally demanding moral theories we considered and abandoned in earlier chapters. If you will recall, I left open the possibility in those chapters that my theory might yet be much more demanding than traditional moral theories have been, or than 'common sense' morality. After all, I argue that our first order theory should have whatever level of demandingness it is that leads to the best consequences in the long run. In that earlier chapter, I defended the claim that this allows for a consequentialist theory that is maximizing on at least one level and yet is not maximally demanding. But I also stressed that this possibility is dependent on contingent factors about the world and human psychology. Given the existence of climate change and other collective action problems, perhaps while it is *possible* for second order consequentialism to allow for less than maximally demanding first order theories, in actuality first order theories should still be maximally demanding, lest they allow for the existence of Polluter's Dilemmas.

But this is not the tack I would want to take. For one thing, all the reasons – the second order consequentialist reasons – we had for our first order theory *not* being maximally

demanding still exist. It is not just unreasonable but counter-productive to be maximally demanding, it does not lead to the best overall consequences. Yes, such a theory would not allow for climate change or food waste or any such thing, but it would also not allow for the existence of options, or for individual projects, or things of that nature. There is perhaps still a bit of a hope that we might find the 'right' level of demandingness, and that because how demanding my theory is depends on contingent facts, the right level would adjust itself to account for global catastrophes like climate change. But I still think that this approach is making a fundamental mistake, and it goes back to the idea of counter-productiveness. Is approaching Polluter's Dilemmas by increasing demands on individuals even the right way to approach the problem, or the way that has the best consequences? It is my suspicion that the answer is no. Instead, we need to move to more collective or group understandings of responsibility.

Moving beyond individualist theories of responsibility

Proponents of a notion of group or collective responsibility often point out that we as a matter of fact *do* ascribe responsibility to groups, or at least talk in ways that sound like we do (Cooper, 1968). We express reactive attitudes to groups as well (Tollefsen, 2003), and otherwise act in many ways similarly to how we act to individuals we believe have committed wrongs. Examples include how we might hold a nation responsible for wrongs it has committed in the past and demand reparations. The best way to make sense of these attitudes and linguistic practices, defenders of collective responsibility argue, is that there exists a notion of collective responsibility. Further, they argue that this notion cannot be made sense of purely on the basis of aggregate individual responsibility because, for instance, the members of a group can change without it affecting the responsibility of the group (Cooper, 1968, p. 260). Overall, one may call

this set of arguments the linguistic or attitudinal argument, since it claims that we already have a notion of collective responsibility in our language and our reactive attitudes.

I am overall less interested in this set of reasons for embracing collective responsibility, though I do acknowledge their force. Even proponents of the view admit, however, that even if we do as a matter of practice ascribe responsibility to groups (nations, clubs, unions, etc.), that does not necessarily mean we are justified in doing so. And as I already said, I think we have particular reason to doubt our intuitions in this matter, so we ought to regard particularly seriously the possibility that our linguistic practices and reactive attitudes about groups are wrongheaded. More to the point, I am interested in notions of collective responsibility not because it is the best explanation for some of our attitudes, but rather because it might provide a way for our moral theories to charge us to solve collective problems like hunger or systemic racism, but without them becoming overly demanding at the individual level in the process.

I think solely individualist ways of thinking about responsibility are inadequate when it comes to dealing with collective action problems, or with systemic problems in social institutions. I am far from the only person who thinks this; indeed, I would characterize this view as one that has grown in popularity in the past few years. Some influential proponents of this view include David Miller (2007), Iris Young (2011), and others (French & Wettstein, 2014). These writers argue that we need a notion of collective responsibility and, accompanying it, a way to charge groups rather than individuals with the moral injunction to solve collective problems. While they generally focus on issues like racism and sexism, global poverty, hunger, and climate change are also often brought up as paradigm examples of the sorts of problems for which the responsibility to solve them lies with groups rather than individuals.

Marion Smiley (2017) calls this *forward-looking* collective responsibility, rather than backward-looking responsibility. The latter is responsibility in the sense of one who is responsible for causing some harm or ill that exists. The former, on the other hand, is in the sense of having responsibility to *do something* about a state of affairs. Of course, the two types of responsibility are related, as most of us believe that being responsible for some state of affairs usually brings with it the moral obligation to solve what problems with that state of affairs may arise. However, they are also separable, and one might be responsible for causing a poor state of affairs without having the duty to make it better, and vice versa. It should not be a surprise that I am more interested in this type of collective responsibility, as consequentialists have always been more friendly to forward looking conceptions of justice and responsibility than backward looking ones. This is because the primary drive for forward looking responsibility is to charge some agent – or group, as the case may be – with bringing out a beneficial state of affairs, rather than assigning blame for past actions. Indeed, my entire project revises traditional notions of blame to be purely forward looking, incorporating backward looking responsibility only if it has instrumental value.

However, the main reason I am interested in this sort of responsibility is that it opens up the possibility of threading the needle between the demandingness objection and Polluter's Dilemmas. A moral theory that charges each one of us with the obligation to deal with climate change, world hunger, or racism cannot help but be too demanding, but a theory that does charge anyone with dealing with these very pressing problems is very clearly inadequate. A hope, therefore, is that we can avoid either by instead placing responsibility for solving these collective ills onto groups.

Act consequentialism and the problem of imperceptible effects

A closely related problem to the Polluter's Dilemma is the problem of imperceptible effects. This problem was highlighted by Shelly Kagan (Kagan S. , 2011) in a paper that started a new debate among consequentialists. The problem of imperceptible effects is that for many of these collective action problems it seems the actions of individuals might be so negligible as to have effectively nonexistent consequences. That is, if one person changes their actions, doing so will not affect the consequences of the collective action. It seems, then, that act consequentialism would be silent about such a change, as the consequences are the same either way. However, if *enough* people change their actions, there will be consequences. And just like the problems of incomparability or aggregation, it is worth noting as Kagan does that this is a problem not just for consequentialism but for any moral theory that cares at least somewhat about consequences, which is almost all of them.

I do want to note that while the problem of imperceptible effects is closely related to the Polluter's Dilemma, it is in fact a different problem. The Polluter's Dilemma is that any moral theory that is not overly demanding does not demand individuals change their behavior enough to avert collective problems. The problem of imperceptible effects is one *reason* for this: that individual actions have such negligible effects by themselves that most theories have little to say about them. However, the problem of Polluter's Dilemmas arises even when actions are not imperceptible, so long as they are morally acceptable. On the flip side the problem of imperceptible effects arises even for maximally demanding theories (and this is how Kagan frames it, as a problem for maximizing consequentialism by its own lights), while such theories escape the Polluter's dilemma but at the price of being unreasonably demanding. A theory will

obviously not allow for unacceptable levels of food waste if it does not allow for *any* amount of food waste, but such a theory is too harsh for most of us.

The same point is made by Julia Nefsky (Nefsky, 2021), who criticizes the general approach developed by Kagan (and others since) which she calls the 'Expected Utility' approach. Kagan's own solution (Kagan S. , 2011) to the problem of imperceptible effects is that individual actions *might* have a morally relevant consequence, because there is a non-zero probability for any action that either it directly has some undesirable effect – the exact molecules of pollution I create *might* find themselves into someone's lungs, for instance – or it is the proverbial straw on the camel's back. The former case does not seem to apply for many cases of collective harms, and the latter very quickly runs into problems of vagueness and sorites-like issues, as it is exceedingly unclear that such a line can even be drawn.

Furthermore, even if the Expected Utility approach does work, it is only a workable solution for maximally demanding theories. That's fine for Kagan, who as noted presents this as a problem for maximizing consequentialists, but unsatisfactory for most of the rest of us. Nefsky, (Nefsky, 2021) for example, diagnoses EU as falling into a dilemma: either EU doesn't explain why we have individual obligations to change our behavior in response to climate change – in which case it allows for unacceptable states of affairs – or it is implausibly strong because it says that even throwing away the slightest bit of food (say) is wrong because that food waste *might* have made a difference somehow. This is, of course, exactly the Polluter's Dilemma I outlined above, and though Nefsky does not use the same term she does note that many other theories fall into the same general dilemma. She diagnoses the problem as being that these theories define the relevant actions (e.g. wasting food, driving a car for leisure) as *pro tanto* wrong. Instead, she argues we should take an 'imperfect' approach, diagnosing actions like these as wrong only if

they are done 'too much'. Nefsky outlines a few different ways to define 'too much', but does not go into great detail. What all of these imperfect approaches have in common, I argue, is that they view these actions not on a case by case basis but *holistically,* in concert with other similar actions either cross-temporally, across multiple individuals, or both.

Similarly, Kai Spiekermann (2014) also critiques some of Kagan's arguments (though he does not wholly dismiss them), and also offers what I think is a more promising solution that does not rely on vague thresholds. Moreover, Spiekermann's solution, unlike Kagan's, can be fruitfully applied to solving the Polluter's Dilemma as well. Spiekermann holds that one of the sources of the problem of imperceptible effects is an impoverished sort of direct act consequentialism that views actions as isolated, discrete events with individual consequences. Instead, we ought to be considering the consequences that actions might have when combined with other actions. Spiekermann remains somewhat skeptical of broader notions of collective actions[33], but still emphasizes that a more holistic view of actions is necessary for solving the problem of imperceptible effects. Rather than judging whether some action might be the one that hits the threshold where consequences are produced, it is better to consider whether an action might have consequences in light of how other actions might also change. Spiekermann suggests that individual actions with minimal consequences might still be wrong if they are expected to bring about harm *together* with other actions. Nor is Spiekermann the only person to suggest that viewing actions in isolation leads to problems.

I brought up the problem of imperceptible effects for two reasons. The first is to emphasize that some sort of theory of how to deal with or incorporate collective actions seems to

[33] To be precise he is skeptical about *Parfit's* notion of collective actions (Parfit, Reasons and Persons, 1987), but Parfit's theory has some peculiarities that do not apply to many other collective action theories which Spiekermann does not discuss in his paper.

be necessary for any theory, because even though maximally demanding theories can escape the Polluter's dilemma they must still deal with the problem of imperceptible effects. While that is one fewer problem, it suggests that the impulse to retreat back to maximal demandingness in the face of the Polluter's Dilemma was a mistaken one. This is especially true because of the second reason I wanted to talk about the problem of imperceptible effects, which is that Spiekermann's holistic view of actions can also be applied fruitfully to the case of the Polluter's Dilemma, meaning that one solution might well do for both problems[34].

Group actions and Holistic views of consequences

Much of my framing of the central issue of this chapter is drawn from the aforementioned paper by Justin Moss, titled 'Strategies for Defusing the Demandingness Objection' (2011). As that title suggests, Moss is principally concerned with how consequentialists might answer the Demandingness Objection. He considers and discards many of the theories we examined in earlier chapters – satisficing consequentialism, Hooker's rule consequentialism, Schefflerian agent-centered theories, and Mulgan's Hybrid theory – though not for the same reasons I do. Instead, Moss is concerned that to the extent that these theories succeed in defusing the Demandingness Objection they fall prey to allowing Polluter's Dilemmas, and vice versa. But Moss does identify one form of consequentialism that he believes can plausibly address both sets of issues, and this is a form of consequentialism proposed by Joseph Mendola (2006) called Multiple Act Consequentialism.

Mendola also criticizes traditional act consequentialism for having a very impoverished view of the 'actions' that it evaluates on a consequentialist basis. Instead, he promotes a notion

[34] Which is not to say that *any* solution to the problem of imperceptible effects also helps with the Polluter's Dilemma – as we discussed, Kagan's (the EU approach) does not.

of *group* actions. Group actions can occur even within a single human life, since many of the actions we take are not standalone choices with no relation to the rest of our lives but rather part of an ongoing project composed of many 'atomic' actions that should not be evaluated in isolation. Similarly, multiple people might work together on a group action – for Mendola, group actions are both an account of actions taken by many people and an account of actions taken over multiple moments in a single person's life. It is a theory intended not only for collective agency but also to explain agency over time.

Moss suggests this conception of actions can in itself allow consequentialists to address both the Demandingness Problem and Polluter's Dilemmas, but I am more interested in using a consequentialist account of group actions in combination with my second-order consequentialist theory of blameworthiness to create a priority ranking for the sorts of actions we should be taking. It is my belief that such a project of prioritization should result in a theory that exhorts us to solve collective problems like climate change in an effective manner, but which still has space for individual failings to a reasonable level and thus does not fall prey to the Demandingness Objection.

Of course, MAC is far from the only theory that distributes responsibility over a group in this way to solve the Demandingness problem, that is the general idea of collective responsibility. I prefer it to other similar approaches for two reasons. Firstly, it is a consequentialist approach, but unlike other collective consequentialist approaches like that of Liam Murphy (2000) it demands that we defect if necessary rather than demand only that we do our fair share. Now, I have said many times that second order consequentialism need not lead to consequentialism at the first order level. However, I also think that we ought to be more like act consequentialists when it comes to large scale cases, since most of the reasons we have for being

nonconsequentialist do not apply in such cases, and collective action problems are typically that[35].

The other reason I am drawn to MAC is that it is presented not merely as an analysis of actions that explains collective actions but also one that explains agency over time. As I said back in chapter 2, I am not fond of modifications to a theory (act consequentialism in this case) that are meant to only solve a particular problem, such as satisficing consequentialism. This is because when one keeps modifying theories in this way the result in the end is likely to have lost many of the advantages that initially drew one to the theory in the first place. Instead, I prefer systemic changes to theories that might have applications in lots of different areas. MAC is an amendment to the concept of 'actions' in act consequentialism that potentially has many different theoretical advantages in different areas, and that is very appealing to me. Thus, overall I agree with Moss that this is a promising direction to take when it comes to collective action problems.

This is particularly because both of these features of MAC are also desirable features for dealing with the problem of imperceptible consequences. The problem of imperceptible acts can also arise over a single person's life as well as over multiple people. Consider the classic example of exercise, for instance: a single session at the gym has such negligible effects that going or not going in any singular occasion makes effectively no difference to your life. Yet regular exercise can make a profound difference to your life. MAC can address this issue in the same way it would address the standard problem of imperceptible effects.

[35] I will present my argument for this in Chapter 11

Spiekermann and Mendola differ quite a bit in the details, and Spiekermann has a much more limited notion of group actions that only groups actions into 'minimally-perceptible subsets' i.e. the smallest group that has perceptible consequences. To continue Moss' analogy, one might say that he believes in 'molecules' of actions but not the unbounded sorts of grouping that MAC embraces. Despite these differences, however, I do think that the main takeaway from both theories is that some sort of theory of collective action, or at least a holistic account that judges actions in combination with other actions rather than in isolation holding everything else constant, is necessary for dealing with the problem of imperceptible effects and Polluter's Dilemmas. Without some notion of group acts, we are much harder pressed to explain why an action might be wrong because it contributes to a larger problem even if it itself might have negligible effects. Since that *is* something we clearly want to say in many situations, I think we need to seriously consider a theory of group or collective actions. While collective action theories have their problems, I consider the difficulties individualist or atomic theories of actions have with imperceptible effects and Polluter's Dilemmas to be a much more pressing concern.

But for all that embracing group actions might solve the Polluter's Dilemma, it seems to impale us right back on the other horn of the larger dilemma. For if any action that has a risk of bad consequences when combined with other actions is wrong, then almost any action seems wrong. At the very least, our theory now seems to have extremely harsh demands on how we live, demanding that we at all times be part of the most beneficent group action we can manage. Moss and Mendola admit that the theory might in fact be extremely demanding in this way (Moss, 2011, p. 128). This is not a problem if you are willing to be maximally demanding, so the problem of imperceptible acts as framed by Kagan (a problem for maximizing act consequentialism by its own lights) might be solved, but for those of us still worried about the

Demandingness Objection we seem to be right back where we started. Group actions are what solve the problem of imperceptible effects and let a maximally demanding theory avoid the Polluter's Dilemma, but to solve the latter *without* being maximally demanding I argue that we need not just a theory of group *actions* but a theory of group or collective *responsibility*.

Moss believes the Demandingness Objection can be elided by relying on the notion of group actions within a single lifetime: in order to effectively participate in such, a certain degree of self-care is required, including not putting undue demands on one's future self (*Ibid.*). This can address some forms of the Demandingness Objection but not, I think, all, and the general direction that one must at all times be participating in the most beneficent available group act, which seems a simple consequence of MAC, is still a very demanding one. Moreover, Moss relies for his solution to the larger dilemma between demandingness and collective action heavily on the following principle: one should 'defect' from a group action - i.e. take an action that is not a furtherance of the group action – only if can achieve better consequences by the defecting act alone than the entire group act achieves (*Ibid.* p. 46).

This principle, without a great deal of qualification, seems too strong and leads to some odd cases. Notably, one can imagine scenarios where the objectively best results would be from everyone defecting but the theory advises that no one defect because no individual can achieve better consequences by defecting (a 'reverse prisoner's dilemma', if you will). Mendola would no doubt reply that if everyone defects they are in fact creating a new group action, but that answer does not work for all cases. None of this is to say the principle is a bad one *per se*, but rather that the necessary qualification does need to be done. The idea of group actions or generally a more holistic view of actions is, I think, a critical component of an answer to the larger dilemma of this chapter, but it is only a component. It solves the problem of imperceptible

effects and can solve the Polluter's Dilemma, but lands us right back into the Demandingness Objection. Which brings us to what I think to be the other component of a possible answer: my own theory.

Second order consequentialism and Polluter's Dilemmas

At this point it is probably helpful to review my method for dealing with the Demandingness Objection. My theory is, in some weak sense, maximally demanding. It says that we ought, for some weak sense of ought, to do the action with the best consequences. However, a certain level of falling short of this ideal, of suboptimal action if you will, is morally acceptable. To be precise, if blaming or chastising you for acting suboptimally would result in bad consequences, my theory says that doing so would be immoral. Since it is highly plausible that an overly demanding theory would result in bad outcomes, my theory allows for some level of suboptimality, satisfying the demandingness objection.

How does this play out once we bring in the idea of collective actions? Let's consider two courses of action available to many of us: reducing our individual consumption to attempt to fight climate change, or participating in larger collective efforts to fight climate change, whether that be voting, campaigning, protesting, etc. Viewed as atomic actions, both these courses of action have a negligible impact on climate change, as one single person's impact on climate change is miniscule. However, if we view these actions holistically, in concert with other actions, the latter is part of a group action with much more impactful consequences. A maximizing consequentialism that had a holistic theory of actions would still demand we do both, but if we had limited time, attention or resources, it would prioritize the latter more impactful action over the former. From a second order consequentialist viewpoint, this means it is plausible that failing to perform the latter action is worthy of blame while not doing the former action, so long as the

consequences remain negligible, may not be blameworthy. Indeed, from the second order consequentialist viewpoint part of the *purpose* of blame and similar reactive attitudes is to exhort people to prioritize correctly. If blaming people for individual patterns of consumption causes them to overprioritize changing that over participating in group actions with much more beneficent outcomes, doing so would in fact be *immoral*.

It is plausible to me that this is so, and also that the earlier concerns we had about theories being too demanding still apply. Thus, the moral theory that SOC would recommend would allow for a certain degree of both selfishness and failing to change individual patterns of consumption and so on. It would still recommend one change one's individual actions *ideally*, but not censure one for failing – in other words, such behavior would be supererogatory. But it would not fall prey to the Polluter's Dilemma, because it is not the case that there are morally acceptable actions that cumulatively result in morally unacceptable outcomes; failing to participate in group actions that fight climate change and its bad effects *is* blameworthy. Further it charges us as collectives to fight these and similar ills, embracing the notion of collective responsibility, and charges us on an individual level with being part of the best collective action we can and prioritizing that over individual contributions that are not part of such collective actions.

It is also plausible to me that this *is* the most effective strategy for fighting climate change and other collective ills,[36] and therefore what SOC would recommend. This is an empirical claim that I am confident in, but nonetheless am much less so than with many of the claims I have made in previous chapters, which were grounded in very basic features of human

[36] Indeed, I think a good case can be made that this is true even over a single individual's life. If I want to alter my future patterns of consumption, for instance, I might well be better off campaigning for laws and policies that would force me to do so than merely trying to alter my day-to-day habits.

psychology. In one sense, I have less problem with this fact than some other theories might. I can be confident that the kind of attitudes and actions my theory suggests we take in response to collective action problems are going to be the ones that are the most effective, simply because that is the criteria my theory uses to determine what attitudes and actions we should take. I don't have nearly the same level of confidence in what those correct attitudes and actions *are*, but if it turns out that my current understanding is wrong my theory will adjust to compensate. I *can* be confident that if there is a way for a theory to not be unreasonably demanding while still avoiding Polluter's Dilemmas, SOC will recommend that. What I have sketched out so far in this chapter is what I think the most promising way of doing that is, and I argue that it relies on a holistic or group theory of action, which in turn leads us to taking seriously the concept of collective responsibility in at least its forward-looking form.

This theory might still seem very demanding. After all, it asks of us that we participate in effective group actions to fight the world's collective ills, of which there are several. Collective responsibility might mean that we need to act together with others, but even if the demandingness of the theory is distributed over several individuals, those individuals still might have to do a lot of work at the individual level. And this is true to an extent, but it is worth remembering that all the reasons we discussed in earlier chapters for our theory not to be overly demanding still apply. Collective responsibility is not my answer to the Demandingness Objection; rather, SOC is. What embracing the ideas of collective responsibility and group action does is change what the moral theory that SOC recommends tells us about where to focus our efforts and prioritize. It does not change the level of demandingness of the theory but rather allows for a theory to be less than maximally demanding without falling prey to the Polluter's Dilemma. Essentially, the theory now says that if you *are* going to be suboptimal, then it is better

to be suboptimal about individual actions and strive for higher optimality with collective actions. It is still a further empirical question about how demanding that theory is, and much as in that previous chapter I still think it likely that said theory will be more demanding than our common sense morality – but it is implausible to me that it would be maximally demanding.

Rather than asking *more* of us, this theory would have us change our notions of what *sorts* of actions are more or less worthy of blame. It will ask us to judge ourselves and each other more on the basis of the policies we support and the group actions we are part of than on our individual conduct with other people on a day to day basis. This is undeniably a change in how we judge people, since I believe we tend to privilege the latter more than the former by default. But if we want to solve the Polluter's Dilemma, and perhaps more importantly the sorts of problems used as examples in that dilemma, I think this is a necessary change.

Problems with collective responsibility

The notion of group actions may once again raise the specter of the Ideal World Objection we discussed in chapter 5, since at first glance that objection rose from the collective nature of rule consequentialism. However, it is not the case that the IWO arose from Collectivization *per se*. As we discussed in that chapter, it rather arises from the fact that, as Podgorski notes, CC makes reference to worlds which differ from ours in more than just our own actions. But the idea of group actions introduced in this chapter does not actually do this – while it asks us to take other people's actions as well as our own into account when judging consequences, those are either the *actual* actions that other people take or the actions we expect them to take given our best available information. Introducing group actions into the theory is meant to change how we calculate the *consequences* of our actions by taking a more holistic approach, but not to collectivize the actions themselves. That is, I am not saying that we should

do that action that *would* result in the best group action if everyone else involved in that group action would also do their part – such a theory would indeed be subject to the distant worlds objection. While we should try to make the group actions we are part of the best they can be, the theory asks us to do so by *actually* changing other people's actions to improve the group action – via campaigning, organizing, and so on – not by making reference to possible words that differ from us in more than our own actions. Thus, incorporating a notion of group actions into our theory does not mean that we are then subject to the IWO.

That said, collective responsibility is not a notion without its own troubles. There are two main general sets of objections to theories of collective responsibility. The first set are metaphysical and conceptual worries about the existence or meaningfulness of group agents or collective agency. The second are second order consequentialist worries about the effectiveness of distributing responsibility to collectives instead of individuals. I am, however, much more worried about the latter set of concerns – not only because I am obviously primarily interested in second order consequentialist implications, but also because the type of collective responsibility I favor has less to fear from metaphysical worries.

The first major set of objections to theories of collective responsibility comes from methodological individualists. These philosophers tend to be skeptical of one or more of a) the existence of collective or group actions, b) the idea that groups can have agency, c) the idea that groups can have intentionality, and d) the idea that something with neither agency nor intentionality can be a proper subject of blame or other reactive attitudes. Of course, no one disagrees that we can talk about the aggregate versions of collective responsibility. Skeptics agree that we can blame groups, but argue that when we are doing so we are merely blaming all the individuals in the group in aggregate, and that there is no further entity that can be blamed.

Proponents of collective responsibility argue that aggregation alone cannot capture what we want out of collective responsibility.

This skeptical worry is not a large concern for me for two reasons. The first is that the type of collective responsibility I am concerned with is the *forward-looking* kind. Forward-looking collective responsibility makes fewer and less weighty metaphysical claims because it is, as Smiley puts it, "not designed to capture an agent's will [but] to distribute moral labor." (Smiley, 2017). The other reason is that I have already developed in earlier chapters a theory of blameworthiness that is heavily revisionary and can accommodate collective responsibility. What it is needed on my theory for us to be able to justifiably assign blame (or other reactive attitudes) is for the target of the blame to be *responsive* to our blame in such a way that our doing so has good consequences. I believe that what makes this true for individuals is that they are reasons-responsive in the right sort of way. The question then is whether groups or organizations can also be sufficiently responsive to moral reasons that our blaming them can have the right sort of consequences, and it seems to me that there are at least some circumstances where this is the case. These two reasons I have for being less worried about metaphysical concerns are related: it is *because* my theory of blameworthiness is ultimately innately forward looking that I am not worried about whether groups can have intentionality or will or other things that individualists worry about.

The more serious set of concerns for me are the consequentialist ones. Many proponents of collective responsibility also use consequentialist arguments, though usually not as their primary reasons for pushing for the concept. Christian List and Philip Pettit, for instance, argue that holding groups like corporations responsible will incentivize them to behave better in the future and encourage members of the group to change their behavior as well (List & Pettit,

2011). But consequentialist concerns against collective responsibility include that it would lead to people being unfairly blamed for the actions of others (Sverdlik, 1987), especially when it comes to group responsibility over a long time (how fair is it to blame people for the actions of their ancestors?). They also include worries that it would be counterproductive, and undermine individual responsibility: this I think is a worry a lot of people will instinctively have about collective responsibility, that assigning blame to corporations and systems lets individuals who had a direct hand in committing bad acts off the hook. Others also think that allowing people to blame problems on the 'system' will let people skate personal responsibility for their own lives.

I share this worry to an extent, as I do not want a theory of moral responsibility that results in no blame accruing to individuals. But that is also true of proponents of collective responsibility, who point out that it is meant to exist alongside individual responsibility, not as a replacement to it (List & Pettit, 2011, pp. 121-122). Assigning blame to groups and charging them with responsibility to better the world does not remove the obligations on us as individuals to do the right thing, though it might change how we ought to exercise that obligation.

My own theory is a combination of collective responsibility with my theory of blameworthiness, and I think it answers some of these objections. The feature I want out of a theory of collective responsibility is a way of prioritizing the consequences of the group acts that we participate in over the individual actions considered in isolation. My theory, in turn, is much more likely to label us as blameworthy for not doing higher priority actions than for not doing lower priority ones. Thus, so long you are participating in beneficent group acts, taking the occasional individual act with subpar but minimal consequences is not blameworthy, satisfying the Demandingness Objection. But *not* participating in beneficent group acts is typically blameworthy, unless you genuinely do not have any opportunity to join more beneficent group

acts. My theory will still recommend that we hold particular participants of a group act as more responsible than others for the consequences of that group act in situations where there is unequal distribution of control over said consequences (so long as doing so would result in better consequences in the long run). The forward-looking nature of my theory also answers the fairness concerns. You are not responsible for the actions of your ancestors, but you *are* blameworthy if you are presently participating in a system designed to benefit you and you are *not* part of group acts to improve that system. The idea isn't to let individuals off the hook for their own actions, but rather to encourage people to prioritize correctly and to consider their actions holistically, as part of larger group act with larger consequences rather than in isolation.

Conclusion

SOC ultimately recommends the theory of collective responsibility that has the best long-term consequences. Unfortunately, I as mentioned am not nearly as confident in what that theory *is* because assessing the long-term consequences of holding groups responsible versus individuals versus some combination of both is no simple matter. However, my current best understanding of said consequences leads me to recommend group or holistic action theories such as MAC. By combining such theories with a prioritization created by SOC, we have a principled way of determining what actions are blameworthy and which do not rise to that level. This allows us to exhort people to solve collective action problems without being too demanding at the individual level. In addition, my theory avoids some of the biggest criticisms about collective responsibility, the metaphysical ones. The literature has been reluctant to get into a full blown consequentialist debate about this subject thus far, but doing so is an approach I think has

promise of being a very fruitful approach that may aid us in answering some of the most thorny problems around collective responsibility[37]. (Smiley, 2017)

[37] Smiley offers a qualified agreement: "Interestingly enough, most of those who offer consequentialist critiques of collective responsibility—and again they are almost always concerned with the practice of holding groups responsible for harm rather than with the facts of responsibility *per se*—do so on a surprisingly general level. In other words, they do not provide us with a set of criteria for thinking about the value of holding groups morally responsible in particular situations. But they could do so very productively on the basis of the more general arguments that Reiff and others provide. Moreover, they could do so without violating their own agent-based approaches to moral responsibility. For, as we have suggested above, being morally responsible and holding others morally responsible are not the same thing. Nor do they have the same relationship to consequences. While consequences may be irrelevant to moral responsibility itself, they may be absolutely key to our choice to hold—or not to hold—agents morally responsible in practice."

PART 4: FIRST ORDER PLURALISM

In this part I will explore the idea of circumstantial moral theories: that a second order moral theory might justify not one first order theory but instead multiple different first order theories that apply to different circumstances. In Chapter 11 I examine the similar idea of threshold deontology and its problems. In Chapter 12 I will consider the idea of thresholds through the lens of SOC, and argue that we both do and should apply different sorts of moral standards to small scale and large scale cases. In Chapter 13 I will apply this pluralism about first order theories to a different kind of problem: that applying our ordinary moral theories to our close personal relationships seems to involve us interacting with those close to us for entirely the wrong reasons.

Chapter 11: Catastrophes, Thresholds and Vagueness

Introduction

It has often been noted that people are more willing to be consequentialist when the numbers get very large. This phenomenon is perhaps most easily – and certainly most commonly – illustrated by means of trolley problems[38]. Few people are willing to push someone onto trolley tracks to stop a trolley that is going to hit five people. But what if the trolley is going to run over a hundred people? What if it is carrying a nuclear bomb that will kill a million people? For most, there is *some* number of people in danger that would cause them to consider the usually unthinkable. This is by no means a universal truth, of course, as there are always absolutists who refuse to compromise their non-consequentialist ethics in any circumstances. Nevertheless, it is common enough to be notable, and for people to try and adapt their ethical systems to account for our intuitions in this matter. In a more general sense, people are more willing to use explicitly consequentialist reasoning when it comes to, say, deciding government policy than in cases of personal moral decisions. It seems to me that this is a different aspect of the same phenomenon, that people are more deontological about small scale cases and more consequentialist about large scale cases. In this chapter, I will look at the prominent nonconsequentialist approach to allowing for different moral principles at different scales – threshold deontology – and argue for its inadequacy. In the next chapter, I will explore the consequentialist view on the matter, and argue that we in fact ought to use different moral principles at different scales, and that SOC can overcome the problems threshold deontologists face in dealing with thresholds and vagueness.

[38]Trolley problems were originally deployed by Philippa Foot as a means to motivate a distinction between doing and allowing (Foot, 1978) but have since been employed by a large number of moral theorists to explore different ethical possibilities, e.g. (Thompson J. , 1985). The discussion here is largely based on Moore. (Moore M. , 1997)

Setting up the problem

The overall picture here suggests that we have two different sorts of moral principles for dealing with two different kinds of cases. The most intuitive way to define these might be by calling them *ordinary* and *catastrophic* cases. A catastrophic case, in this way of defining the difference, is a case where following the constraints that a nonconsequentialist theory imposes will result in bad consequences at much more extreme scales than in ordinary cases: consequences like the deaths of thousands or widespread pain and misery across an entire population. The idea that nonconsequentialist moral theories, and especially rule-based moral theories like deontology, might have exceptions to deal with catastrophic cases is quite an old one. Nearly every rule-based theorist builds into their theory some sort of exception for moral catastrophes. Not all deontologists follow this tack, though, with some, such as Anscombe (Mr. Truman's Degree, 1956) (1958, p. 17), not allowing for the relaxation of their constraints even if the alternative is the deaths of many people or some other catastrophic consequences; we will call this the *absolutist* response.

If one is not willing to be an absolutist, the main alternative way of adapting deontological theories in the face of moral catastrophes is by means of thresholds. A threshold deontologist or threshold non-consequentialist believes that above a given threshold of possible harms that might be prevented, usual non-consequentialist constraints (such as the constraint against murder, or lying, or doing harm) are relaxed. Such a view can of course be refined in various ways, such as adding differing thresholds to different constraints so that lying to save a single life is acceptable while murder is only acceptable if it saves many thousands. But all such threshold theories share a number of serious concerns. The main set of concerns affects the arbitrariness of a threshold, as such theories have a hard time justifying any particular number or

even a range of numbers in a manner that does not seem *ad hoc* or lead to the theory collapsing into a form of consequentialism.

The problem of numbers

Despite the strong feeling that ordinarily impermissible actions might become at least permissible if not obligatory to prevent truly catastrophic consequences, absolutism does have its defenders. One way in which a deontologist might make denying the existence or force of moral catastrophes more palatable is to emphasize the problem of aggregation. Many deontologists think that harms cannot be aggregated or at least cannot be aggregated simply, that a million deaths are not actually a million times worse than one death. John Taurek (1977) takes the argument that numbers do not matter to its logical but also highly troubling conclusion: that numbers do not matter at all. I have never found Taurek's arguments to be even slightly convincing even leaving out his radical conclusions. But I also feel the need to come to Taurek's defense in this, as the same could be said for the other side. For me the claim that it is better to save the many than the few is so basic that it is genuinely difficult to argue *for* it, but my intuitions in this regard seem to be something Taurek does not share, just as I seem to not share the intuitions he uses to support his arguments.

Taurek starts arguing for his claim by an example: we would think someone to not be acting wrongly if they choose to save themselves rather than five strangers, if the means by which they may save either themselves or those others is theirs to distribute. Immediately, several responses become available, the most immediate of which is that just because the person might not be morally required to save the five over the one in that instance, it does not necessarily follow that the numbers don't count. Even if we leave aside my theory of blameworthiness, I think in situations like this it is more plausible that the person saving

themself has a good enough *excuse* that their failure to save the strangers is permissible rather than that they genuinely have no duty to save the five over the one – that is, that they have an *overridden* reason rather than *no* reason. Taurek simply denies the intuition that we have a *ceteris paribus* reason to save the many over the few that can be sometimes overridden by other considerations (p. 303). He bases this denial in large part on the claim that the impersonal point of view is impossible to grasp and cannot motivate actions, but I actually find this denial of an overridden reason for saving the strangers to be more interesting, and the more difficult to accept. Once Taurek expands the example to third parties, he loses me completely, as my intuition about such cases is the opposite of his.

That said this example is used mainly as a springboard, as much of his argument against numbers is based on the aforementioned skepticism about an impersonal point of view. However, I don't actually think the impersonal point of view is necessary to feel the weight of numbers – something like an intersubjective point of view (i.e., seeing things from the view of multiple people at once) seems to suffice to explain why it is better to save the many over the few. And it would be, I think, less troubling to someone like Taurek who is skeptical that states of affairs can be good or bad *simpliciter*. One can make a reasonable argument that something cannot be good unless it is good *for* someone, but I think it's much harder to say that something cannot be good for a *group* of people, or indeed for people generally. Further, I can empathize with groups as well as individuals, or so it seems to me. And so a lot of Taurek's claims rest on what I view as rather shaky foundations.

Taurek's final and most persuasive argument (for me at least) is an argument from fairness that asserts that treating humans as equally valuable to each other means giving no weight to numbers. He argues that having equal regard for each individual person entails that in a

situation where we must choose between saving the one or the many the right thing to do is to flip a coin, as that would give everyone an equal chance of survival. Because of the difficulty with engaging with Taurek's premises at any level other than blunt intuition sparring, much of the literature around Taurek does not directly engage with them. Instead, it usually focuses on showing how his moral principles and other similar lottery-like principles lead to contradictions (Choen, 2014) or undesirable results (Bradley, 2009). While I think these arguments are successful, it is ultimately that intuition sparring that is the true core of Taurek's argument.

For my part, I will once again hearken back to earlier chapters and say that I think the intuitions that Taurek relies on are among the most unreliable and suspicious of our intuitions and we have good reason to doubt them. This is especially true of the idea that we do no wrong when we decide to distribute something that is 'ours' in a way that privileges ourselves. This is an intuition that I think is used to justify too many real and present inequities for me to be comfortable accepting it uncritically. Similarly, if we are not willing to accept that states of affairs can be better or worse for groups and not just for individuals, we will have a much harder time dealing with collective action problems and deciding large scale policies. Such arguments may seem to be question-begging: someone skeptical of aggregation and other foundational premises of consequentialism is unlikely to be impressed by my second order consequentialist argument against the moral principles that Taurek relies on. But I think it becomes less circular – or at least less *troublingly* circular – when we view our moral theories holistically, as entire systems that stand and fall on their sum total merits rather than on individual logical failures. That's also why I put less emphasis on deriving contradictions from Taurek's chance-based moral theory, though I must also note that, like Bradley, the real-world implications of adopting such a theory are unacceptable to me.

Having said all that, not everyone that follows in Taurek's footsteps thinks that numbers don't count at all: many take a more moderate position that Choen (2014) calls 'Numbers Partly Count". That is, while most deontologists are at least skeptical of *straightforward* aggregation (i.e. that the death of five people is exactly five times as bad than the death of one person) they are more willing to accept that the deaths of many might be worse than the deaths of a few, which still leaves the door open for moral catastrophes. Nozick, for instance, is one of the foremost and most strident proponents of the importance of the separateness of persons and the claim that harms cannot be simply added together, but still acknowledges the existence of moral catastrophes (Nozick R. , 1974, p. 30). Other examples include Sanders (1988) and Lawlor (Lawlor, 2006) who explicitly defend a view that hews close to Taurek in cases like one vs. five but departs from him in cases like one vs a million. In other words, they embrace a form of threshold deontology, meaning they also inherit the problems of thresholds which, as we shall see, are considerable. Still, this shows that the alternative of absolutism remains unpalatable even to many of those who are skeptical of aggregation.

Going without thresholds

On the other hand, one can acknowledge that numbers at least somewhat count but still be an absolutist: holding, for instance, that in a choice between saving one and saving five one should do the latter, but that, say, murder is impermissible even if the alternative is the death of thousands. Kant himself once famously said that rather than compromise on his moral theory it was "better the whole people should perish" (1780). While few deontologists are willing to be as bold as Kant about sticking by their theory even if the alternative is the ultimate catastrophe, there are other arguments about why moral catastrophes aren't really a problem for deontological theories that do not involve bringing in the idea of thresholds. However, as Larry Alexander and

Michael Moore note, these responses all have some flavor of evasion from the deontologist (2020). G.E.M. Anscombe's argument that a truly moral agent will never contemplate such cases and so will never be presented with such a dilemma (Anscombe, 1958), for instance, is merely a peculiar form of absolutism; one that, moreover, makes the view less palatable rather than more, as it amounts to an abdication of the question.

Various other responses also abdicate the question in this way, like Bernard Williams' assertion that cases like this are beyond morality or moral reason (1973). Such a view seems at first glance simply pessimistic of morality's scope, but I would argue that just as Anscombe's response is absolutism with an additional, as Alexander and Moore put it, impatience with the question, Williams' response is threshold deontology with the same impatience. I say this because Williams *does* seem to think that in such situations we would do what we have to do (e.g. kill one to save thousands), he just thinks the reasoning behind our actions is not worthy of being graced with the name of 'moral reasoning'. But, in the end, that is a dispute more about nomenclature than about the structure of our approach: Williams still thinks that we ought to apply one type of reasoning below the threshold and another type above it. Naturally as a consequentialist I deny that the reasoning I use to evaluate these sorts of cases somehow doesn't count as 'moral', but a lot of deontological responses along similar lines (distinguishing moral reasons from all-things-considered reasons, for instance) also seem similarly to be a variant on threshold deontology. Certainly, they share the same flaws.

If most alternatives to absolutism or threshold deontology are variants on the basic structure of the latter (one type of reasoning in ordinary cases, another in catastrophic cases) or amount to evading or abdicating the question, are those the only two real options for the deontologist? I think the answer is yes and no: that is, I think there is a third option, but one

could argue that it is yet again a variant of threshold deontology. The reader will likely be unsurprised at what I think the third option available to the deontologist is, and that is a second order deontology. I will argue below that a second order theory could allow the deontologist to avoid many of the problems that plague threshold deontology, at least in principle.

The other main framework that attempts to account for intuitions about differing moral principles at different scales is via indirect consequentialism. Indirect consequentialism approaches the problem from the other direction, as it were. Rather than trying to explain why we should give up on (or at least relax) non-consequentialist considerations when the numbers get very large, indirect consequentialists come up with arguments for why we should, for instance, *not* kill one person to save five in cases like the Fat Man or Transplant. That is, indirect consequentialists try to explain our intuitions that we *ought* to obey constraints such as 'don't kill' in all but exceptional circumstances in consequentialist terms. Indirect consequentialism avoids some of the problems of threshold deontology, but of course also comes with its own set of problems. I believe, however, that by moving from indirect to fully SOC, many of those problems can be avoided. In this chapter, I will discuss why I do not think that threshold deontology is a good framework for grappling with the problem of moral catastrophes or even just large-scale cases in general. In the next chapter I will discuss consequentialist alternatives, including my own.

The problem with thresholds

As noted above, while some deontologists are absolutists who hold that no amount of adverse consequences trump deontological considerations, even strict deontologists sometimes allow that moral rules can be overridden if necessary to avoid sufficiently catastrophic consequences. The basic idea behind threshold deontology is that deontological norms govern up

to a point despite bad consequences for adhering to them, but that at some point the consequences become so dire that the norms must be set aside. These theories can be elaborated on in various ways, where for example the threshold of adverse consequences above which telling a lie is justified might be high but still lower than the threshold for murder, for example. Michael Moore (1997), in particular, develops this idea into a fully developed theory of threshold deontology and his theory shall be the main example I consider in this chapter. Threshold deontology can be thought of as an attempt to modify regular deontological theories to account for the fact that we seem intuitively willing to ignore moral prohibitions for large enough good consequences or to avoid large enough bad ones, and also to avoid the charge of deontological theories being too fanatical about following rules.

Threshold deontology has a problem, however, with specifying *where* the threshold lies. The idea that Moore seems to have in mind is as follows: deontological constraints apply until the number of lives at risk reaches some number N, and after that purely consequentialist considerations dominate. Moore does not specify the precise N, holding that doing so is unnecessary just as defining the precise number of grains of sand in one place that become a heap is unnecessary. That is, Moore does not provide a solution to the philosophical problem of vagueness, but rather argues that threshold deontology does not produce a *special* vagueness problem that is more dire than the general one, so Moore can avail himself ultimately of whatever philosophical solutions to vagueness seem the most plausible. As he puts it: “Although this worry [that there is no way to judge the exact threshold] can surely give rise to quite genuine perplexity and anxiety when we make practical decisions, it is not the basis of any very powerful objection to threshold deontology as a moral theory.” (*Ibid.* p. 754). However, I don’t think that Moore can claim this so easily. I think there are two worries about thresholds that Moore here

conflates. The first is that we don't seem to have any way of judging where the threshold lies, to which Moore responds that the threshold is vague. The second, more serious worry is that we don't know *why* the threshold lies there – for this worry, appealing to vagueness does not actually help.

At this point it is probably helpful to outline the general problem of vagueness. As I alluded to earlier, vagueness is generally analyzed via the problem of the heap, the sorites paradox first put forth by ancient Megarian philosopher Eubulides (*soros* is the ancient Greek for 'heap'). The basic problem of the sorites paradox is that it defies normal inductive logic: a single grain of sand is not a heap, if you have some number of grains of sand that are not a heap then adding a single grain of sand to it will not make it a heap, by mathematical induction there is no number of grains of sand that make a heap. But obviously if you put enough grains of sand together, you get a heap. In recent decades, there has been a renewed interest in the problem of vagueness, with the main point of contention being what sort of logic, if any, applies to cases like these: basically, the question philosophers try to ask is what sort of approach do we or should we take to sorites cases. I don't want to get too deep into the literature on vagueness, fascinating though it is, but rather I want to confine myself to the previous assertion of Moore's. I would like to argue that appealing to vagueness doesn't solve the problem that we do not have a way of telling where the threshold lies.

This point was first brought up by Anthony Ellis (1992), who compares moral thresholds to the one between red and orange:

> Since the response is such a popular one, the point may bear repetition. Take an analogy. There is no precise cutoff between, say, red and orange, and if we wanted one we should have to specify it arbitrarily. But we could not put it just anywhere within the color spectrum. We should be faced with specifying an arbitrary point, but within a non-arbitrary range. That is not what we are faced with in morality. Specifying a range where

> the transition from wrong to right takes place would be no less arbitrary than would be specifying a precise cutoff. […] The concession that the range is a vague range makes it no less arbitrary. (p. 869)

Ellis goes on to note that in the case of red-orange, we can appeal to the broad agreement in judgements that people have about the color spectrum to identify the range, even if there is not agreement about the precise cutoff point. But this is something we cannot do in morality:

> It is not just that we do not find agreement in judgements here. (Though we don't, and this is to be expected on my account.) If that were all that it is, then we could simply conclude that, where there is not such agreement, the matter is indeterminate, not yet decided. The problem is that in morality such judgements cannot be settled by group agreement. […] Suppose that everyone did in fact agree, on the number 50 say, *but no-one could give any reason why it should be 50 and not some other number*. This would not tell us anything about moral theory; it would simply be an utterly bizarre mystery. And if we found agreement on a vague range—again with no-one able to give any reason to justify this range rather than some other—then this might be less dramatic but it would be just as mysterious. It would be senseless to think that this agreement could make any contribution to settling the moral question. (*Ibid.* pp. 869-870)

What Ellis is getting at here is that moral theories cannot simply appeal to people's judgements about a case, nor can they appeal to bare intuitions. Even if we could identify the range in which the vague threshold exists (already dubious), the threshold deontologist still has to explain *why* the threshold is there. Moore acknowledges that he cannot provide a non-arbitrary reason for the threshold to be at some precise N, and so responds by appealing to vagueness and saying that there is no precise N. But this doesn't actually provide a non-arbitrary reason for the threshold. The problem is not that any *particular* N is arbitrary, but that *any* particular N is arbitrary.

All of this is not to say that there are no theories of vagueness that Moore can incorporate into his theory. Vagueness is still a young and burgeoning field in philosophy at the time of this writing, and it is possible that there are some ways of understanding vague statements that Moore can embrace without much trouble. What it *is* to say is that Moore cannot simply say that the problem of vagueness that his theory has is 'simply the same problem of the heap'. Vagueness in

the moral realms is more problematic than the basic sorites paradox, for reasons that should by now be clear. Simply put, most of the literature around vagueness attempts to analyze ordinary language terms that are vague, like ‘heap’ or ‘red’. But we hold our moral theories to a much higher standard than our ordinary linguistic practices, or at least most of us do. It is one thing if your theory of vagueness has the dividing line between ‘red’ and ‘not red’ turn out to be essentially arbitrary, simply a matter of group agreement, or even a linguistic or psychological illusion. It is quite another thing entirely if your moral theory has the difference between *right* and *wrong* turn out to be essentially arbitrary or an illusion. There is an additional problem of *justification* that is not there for heaps, because moral claims have a different sort of justification than claims about heaps do. If anything, vagueness turns out to be a bit of a red herring here, though much of the debate around threshold deontology has revolved around it. The real problem was the one of justification all along.

Indeed, there’s an argument to be made that precise thresholds weren’t a problem in the first place. In a direct response to Moore’s theory, Larry Alexander (2000) provides several examples of cases where a precise threshold leads to strange and unsatisfying results, which largely involve the number of lives in danger suddenly dropping above or below the threshold and how the subsequent radical change in proscribed outcomes does not jive with our intuitions. For instance, let us say that torture is permissible to save N people, so the police capture and torture a terrorist’s mother to save the N people who are being held hostage with bombs. The terrorist agrees to let one of them go. But now, there are no longer N people in danger, but N-1 people, so torture is no longer permissible. It appears, thus, that threshold deontology has allowed for a situation where torture is permissible even though it saves only one person – something even consequentialist theories generally do not do – so long as that one person is the

crucial person on the threshold. We can also imagine the opposite situation, where there are too few people for torture to be justified, but if the police send in *more* people to be hostages, they will then be justified in using torture. Even a threshold that consists of a range rather than a precise number will run into this kind of problem if the range is small – and if the range is too large, it seems like we would have a different problem of our moral theory being *too* unspecified.

But the threshold deontologist might well respond that this sort of sudden jump from some action being permitted to being forbidden isn't actually that odd, and in fact it happens all the time with other moral theories. For example, we can observe, as Richard Arneson (2018) does, that a consequentialist might also say that torture is unjustified to save N people but justified if N+X people are in danger – because in the former case the negative consequences outweigh the positive ones and in the latter case the reverse is true. One can refine examples until X is small, and how is this any stranger than the case of the threshold deontologist? Arneson is, I think, correct that this sort of sudden jump isn't necessarily strange or incoherent, and Alexander acknowledges in his response to Arneson (Alexander L. , 2018) that his examples are perhaps lacking. But what Arneson misses here is that what those examples are trying to highlight is less the discontinuity in *outcomes* so much as the gap in *justification*. In the case of the consequentialist, there is no mystery in *why* the action goes from forbidden to permitted, as it falls out straightforwardly from the consequentialist's theory of the good. Lacking the ability to appeal to such an argument, the threshold deontologist does have a problem of explaining what *causes* the sudden jump from forbidden to permitted. Here is Alexander's full response:

> Arneson correctly notes that some of my weirdness examples fall short. […] But my main skeptical point in that article was over the absence of any theory to justify the location of the threshold. Consequentialism has a theoretical basis. So, too, does deontology. Neither rests solely on intuitions. But is there, and can there be, a theoretical basis for the location of the deontological threshold? If Al intuits that the fat man can

> only be permissibly shoved in front of the trolley to save five lives, Sally intuits that shoving him is permissible only if ten lives are at risk, and Dana thinks the threshold is twenty, is there any theory that can establish whose intuition is correct? (Alexander L. , 2018, p. 435)

Ellis also challenges Moore on a related point, noting that the problem is not just in locating the threshold but in justifying the *existence* of the threshold. Deontic claims might have exceptions, but these exceptions must be justified. Ellis notes that Moore cannot justify the threshold's existence *merely* on the basis that the number of lives lost is large, since such consequentialist considerations presumably applied *below* the threshold as well. Thus, those very same considerations cannot be used as the justification for why those considerations start to matter when before they did not matter. Moore seems to think that there is intuitively a point at which consequentialist considerations outweigh deontological considerations, but Ellis rejoins that this assumes that deontological considerations and consequentialist considerations are commensurable, an assumption Ellis rejects.

Another way of putting the same point – and another way to respond to Arneson's points from earlier – is to note that any attempt to create some sort of commensurability between deontological and consequentialist considerations seems hard to distinguish from some sort of weighted consequentialism. That is, a consequentialism that assigns different 'weights' to the goodness or badness of consequences depending on, as one example, the distinction between killing and letting die. For instance, one might say that killing someone is impermissible unless it would save a million or more people (and killing two people is impermissible except to save two million or more people, and so on). But this claim *seems* like a claim that killing someone is as bad as letting a million people die in the consequentialist calculus, but that we are otherwise using said calculus. In fact, it can be shown that threshold deontology with a sliding scale is extensionally equivalent to an agency-weighted consequentialism (Sen, 1982).

A weighted consequentialist might defend such a weighting on similar grounds that a deontologist might motivate a prohibition against murder – e.g. holding that we are constitutively bound to respect human lives – or via further consequentialist reasoning – e.g. holding that prohibiting murder save in extraordinary circumstances leads to the best consequences in the long run. But either way, they alter the consequentialist calculus without claiming that there is any point where it does not apply. By contrast, a deontologist is, by presumption, someone who rejects that such a calculus is the correct criterion for moral rightness (or at least, in the case of the threshold deontologist, below a certain level); Moore certainly seems to want to resist the idea that allowing consequentialist criteria to override deontological ones on occasion amounts to some sort of consequentialism[39]. To the extent that he wants to avoid such a collapse, therefore, he must resist allowing consequentialist considerations and deontological considerations to be placed on the same scale. But if he does, there seems to be no escape from Ellis' charge of arbitrariness.

Alexander (2000) expands the argument against threshold deontology in a similar direction to Ellis, also coming to the conclusion that a consequentialist has better grounds to achieve the same result of differing moral principles for small scale and large scale cases (though he is not himself a consequentialist). The reasons for which a consequentialist might advise sticking to absolute prohibitions save in exceptional cases have already been gone over in previous chapters[40]: decision costs, cognitive limits and biases, coordinative difficulties, etc. Further, Alexander notes that there are reasons a consequentialist might want the threshold to be

[39] Moore argues that his theory does not amount to weighted consequentialism because a consequentialist who thinks murder is worse than letting die is still obliged to murder in order to prevent five murders. It is possible, though, to incorporate a doing/allowing distinction as a weight as well, so this doesn't preempt the collapse objection as well as he hopes.

[40] Chapter 4, pp.50-52

unspecified rather than precise. Firstly, there are all the problems he raises with precise thresholds that we have already discussed. Secondly, there is the danger that specifying a precise number leads to people being too ready to apply strict consequentialist reasoning rather than the heuristics that are actually more likely to lead to better outcomes, as has been mentioned earlier. Thus, the indirect consequentialist can provide an explanation for why there is a threshold, where that threshold is, and why it ought to be vague, all by appealing to their theory of the good and the basic consequentialist calculus. The threshold deontologist lacks any similar explanation.

So, while threshold deontology and indirect consequentialism both allow for differing heuristics for ordinary and catastrophic cases, the threshold deontologist struggles to account for the existence of such a threshold in a non-arbitrary way without collapsing into a weighted consequentialism. By contrast, the indirect consequentialist does not have the same problem of an arbitrary boundary, because while they promote differing moral principles for differing cases the *justification* for those principles uses the same reasoning, and that reasoning is applied smoothly throughout all cases. However, that does not mean that an indirect consequentialist can totally escape the problem of dealing with vague or borderline cases in its entirety, as we will discuss in the next chapter.

The defense for threshold deontology

So, does this mean that I think deontological theories should be disregarded on the basis that they cannot come up with a satisfactory non arbitrary justification for thresholds that does not collapse into weighted consequentialism? Well, not quite, or at least not entirely, for a few different reasons. I do think that deontologists have to either be absolutists or embrace something structurally similar to threshold deontology, allowing for different sorts of moral principles to apply in ordinary and catastrophic cases. I also think that absolutism is untenable, and that

threshold deontology cannot work for the reasons outlined in the previous section. But that does not mean that deontology cannot address these problems.

The reader probably remembers when I was on the other side of a similar 'collapse' objection. Indeed, when I mentioned that threshold deontology was extensionally equivalent to weighted consequentialism, one might recall that I spent much of chapter 4 arguing that even when two theories are extensionally equivalent (in that case, indirect and second order consequentialism) we still might have good reasons to prefer one over the other. Granted, this argument does not seem to serve as well for threshold deontology, since in this case even the purely theoretical advantages also seem to be on the opposite side. A consequentialist approach to thresholds holds purely theoretical advantages over threshold deontology for the reasons Alexander describes – because it provides a justification for the existence of a threshold and its vagueness whereas threshold deontology does not - so given extensional equivalence we have reason to prefer it.

But the problem for the deontologist, I think, lies in trying to approach the problem from the side of *first order* ethics. The threshold deontologist's main problem, as Ellis expounds on at length, is in justifying the existence of a threshold, justifying why there is a discontinuity where consequentialist considerations are allowed to overcome deontological constraints when they were not before: both why at that point in particular and why at all. This sort of justification is precisely the business of *second* order theories, not first order ones, and Ellis rightly judges that a first order deontological theory doesn't have the right tools to give a proper justification of this type. In Moore's response to Alexander's criticisms of his theory, we can see this focus on purely the first order side. Moore develops a notion of stringency, outlining the framework of a threshold deontology where less stringent deontological obligations are overridden by more

stringent consequentialist ones. But even he admits that he has not done the work of *justifying* the threshold. (Moore M. , 2018, p. 387) All he really does is outline what different plausible threshold deontological theories might look like: hardly worthless work, but very much a restatement and clarification of the problem rather than a solution. As Alexander points out (2018, p. 438), it *sharpens* the problem if anything: Moore provides several intuitively plausible ways to denote a threshold, but provides no way, nor even the outline of a way, that we can decide *between* them. Alexander doesn't use the same terminology as I do, but it is clear to me that he and I have the same criticism here: threshold deontologists need to provide a second order theory that justifies thresholds and not merely a first order theory *of* thresholds, and thus far have not done so.

That said, part of the reason it seems like consequentialism has an upper hand here, I think, is that consequentialists are much more accustomed to thinking in second order or near second order terms. Because the straightforward application of consequentialist metrics to decision making in ordinary cases leads to unintuitive and unappealing results, consequentialists are used to applying their theory indirectly, and have been doing it since the very first utilitarians. Second order consequentialists like me can draw upon a long history of indirect consequentialists, two level consequentialists like Hare, or rule consequentialists like Hooker, all of whom we have discussed in previous chapters and will discuss more in the next. By contrast, deontological theories when applied directly lead to unappealing results in *un*ordinary cases – what we have so far been calling 'catastrophic' cases, though I will argue in the next chapter that they are more common than that label makes them sound – and so deontologists generally don't think in indirect terms. Therefore, they simply do not have the same well-trodden ground to draw on. But much as with the case of Demandingness, I think that the case that some sort of second

order theory is needed to properly address the problem of thresholds is stronger than the case for second order *consequentialism* in particular, though it is much more obvious here than there that consequentialists have a significant advantage (at least at present).

A sufficiently detailed second order theory will provide us with a justification for why thresholds exist, possibly a justification of why we should keep them vague (as Alexander provides for SOC), and an outline of how to weigh our different sorts of moral systems against each other and what to do when they clash. That last point might also seem to speak in favor of consequentialism at the second order level, as the very language of weighing and clashing seems to presuppose commensurability: Ellis' criticism. And it has to be said that this is also a large part of the reason why consequentialism of either the indirect or second order variety has a great advantage here. If you think that different sorts of considerations cannot be weighed on the same scale, it is much harder to explain how one could outweigh the other – and especially hard to justify the existence of a *particular threshold* where one will come to outweigh the other. However, at least in principle a sufficiently specified set of norms with a detailed list of priority rules and exception clauses can do the same job as a weighted consequentialism (Richardson, 1990), even if such an approach may seem implausibly convoluted to most.

Which brings us to the final and main reason I don't think that the problem of thresholds is the death knell for deontologists, and that is because I do not believe in single grand arguments that can invalidate an entire approach. I am, on the second order level, a consequentialist, but that is not because I think that nonconsequentialists have some particular case that their theories cannot grapple with. Rather, it is because when viewed *on the whole* I find the consequentialist approach to have more advantages and fewer disadvantages when compared to the nonconsequentialist approach (at least on the second order level). Although I find it presently

unlikely, some future second order nonconsequentialism may yet disabuse me of this belief: as I said earlier, consequentialism does have something of an advantage here in there being a long history of indirect theories to draw from that nonconsequentialists don't have. In a similar vein, I do not advise my reader to from this chapter alone conclude that threshold deontology is entirely unable to deal with the problems of justification and vagueness, but to consider that the fact that indirect and second order consequentialism have an easier time with them as a reason to investigate those approaches further to see if they yield more similarly fruitful results in related areas, as I believe they do.

Conclusion

Deontologists are sometimes accused of 'rule worship', of hewing to their principles regardless of the cost: "though the world perish", as it is sometimes memorably put. Threshold deontology is an attempt to soften this critique by allowing for exceptions to deontological rules in certain catastrophic circumstances. However, threshold deontologists suffer from a problem of arbitrariness, as they fail to provide a justification for this exception, or for why the threshold should lie where their theory says instead of some other place. I argued that the source of this problem is that they only provide first order theories, but that this sort of justification is the business of second order theories. SOC could allow a threshold deontologist to escape this problem, but so could any other second order theory. Such an alternative to SOC would need to be developed, however, and so far has not been. That said, consequentialists do have a head start in this regard due to a long history of indirect theories. Still, I will not take on the task of developing such a theory for them. In the next chapter, I will continue to apply SOC to the cases of thresholds and scale.

Chapter 12: Circumstantial Moral Principles

Introduction

In the last chapter I argued, following in the footsteps of Ellis and Alexander, that threshold deontology faces a deep arbitrariness problem. Ellis charges that the source of this arbitrariness problem comes from the fact that deontological constraints and the consequentialist weight of numbers are incommensurable considerations, or at least that they must be if deontology is not simply to collapse into some sort of weighted consequentialism. However, I suggested that this problem is in fact due to threshold deontology lacking a proper justification for the threshold, and may be resolvable with a second order theory: in this chapter I will try to do so with second order consequentialism. Further, I will argue that this problem of incomparables can be generalized beyond the case of thresholds. I hold that we have some very different and seemingly contradictory intuitions about different sorts of cases, most notably what I shall call 'small-scale' and 'large-scale' cases. The moral principles we apply to these different sorts of cases do not seem to be consistent with each other. However, I will go on to argue that this seeming inconsistency can be not just resolved but justified: that not only do we hold different and incompatible moral principles that we apply do different sorts of cases, but that this is in fact what we should do. The presence of a second order theory that underlies these disparate sets of first order action-guiding principles prevents this from generating irresolvable contradictions.

The consequentialist solution to thresholds

In the last chapter I approached the issue from the angle that deontologists will approach it, where the constraints they normally believe apply to our actions – such as the prohibition

against murder – seem to relax when the negative consequences of abiding by them becomes very large. Or, at least, that is the feeling that most of us have about catastrophic cases, and even many ardent deontologists like Nozick are unwilling to bite the absolutist bullet.

Consequentialists, however, have to explain the opposite phenomenon: namely, why it is that in most ordinary cases we have the strong sense that we *do* need to abide by such constraints. By this point we have discussed indirect consequentialism enough that the reader is familiar with the many arguments consequentialists use to argue for this: the decision-cost of making a calculation for every decision versus following well considered rules of thumb; the possibility of unconsidered and unintended consequences, especially over the long-term; the fact that we often operate under uncertainty such that we cannot fully trust our evaluation of the consequences and as such should err on the side of minimizing the direct harm we cause; the fact that we need to consider what courses of action we legitimize in similar circumstances in the future; and so on and so forth. Conversely, in catastrophic cases many of these considerations do not apply. The decision cost is relatively lower and the negative consequences of abiding by the constraint are far more serious, lessening the likelihood of the long term consequences of following it being better. In addition, the worry about precedents is less dire, as can be seen clearly by the name we choose to use in cases like this: ‘catastrophic’. We already intuitively delineate these kinds of cases so that they do not reflect back onto ordinary situations, unlike cases such as Transplant. In agreement with Alexander, I think indirect consequentialists have a much easier time with classic threshold cases such as we discussed in the last chapter, for reasons we also discussed there.

As also discussed in earlier chapters, indirect consequentialists are not rule consequentialists. They do agree that following rules of thumb that are similar to what a rule

consequentialist would advise is the best course of action in many ordinary cases, but if those cases are clarified to remove the various factors that make them so agree – factors such as incomplete information or the awareness of long-term consequences – they do not recommend that one should still follow those rules. While I have a fair amount of sympathy with the argument I mentioned then that our intuitions are unreliable in unrealistic cases, nonetheless I believe that consequentialists have the resources to argue that we should continue to follow these rules in cases like Transplant. We went over these arguments in detail in chapter 4, but to quickly recap, I think that for some rules to be effective at generating the best consequences in the long-term we need to bind ourselves to them and internalize them, not merely adopt them as rules of thumb to be discarded when convenient.

Yet I do not also advocate for simple rule consequentialism, because I believe that in certain other kinds of case we should be more straightforward act consequentialists. Rather, I hold that all the normal arguments for indirect consequentialism amount to a *justification* for why we should be rule consequentialists in ordinary cases, while we should be act consequentialists for other cases. Since my schema amounts to having different sets of rules for judging different kinds of cases, it is despite being a form of consequentialism at the second order level closest of all the discussed approaches to threshold deontology *at the first order level*: it advocates that below a certain threshold of bad consequences the criteria for judging the rightness or wrongness of actions is moral rules, but that above that threshold the criteria we should use is solely the consequences of the action.

Yet, because it *is* ultimately a form of consequentialism, I believe that it avoids the pitfalls of threshold deontology discussed above. There is no odd discontinuity in my approach, because the *underlying* justification for why we should adopt the moral heuristic we should is the

same even if said justification recommends different schema for different circumstances. Further, the charge of arbitrariness is less worrying for the same reason, since at least in principle a fully fleshed out version of my theory will spell out exactly when we should apply different sets of moral rules. Even the problem of vagueness is mitigated, because as Alexander argues there is a consequentialist justification for why the threshold *should* be vague rather than precise (since an overly precise threshold will lead to undesirable consequences). As trolley problems often demonstrate and threshold deontologists attempt to explain, our intuitions do seem to be more consequentialist when large numbers of people are involved. I do not think it therefore follows that we should not stick to rigid constraints in other kinds of cases, because I think we are justified in following different moral rules in different kinds of cases.

These are the benefits of my theory over threshold deontology, but one might well also ask what the benefits of such a theory are over indirect consequentialism. To some extent, this is a question I already considered and answered in chapter 4, and at least some of the advantages are purely theoretical with little practical difference. I also believe that second order consequentialism has many other advantages over indirect consequentialism in other areas, which we have discussed in earlier chapters. But when it comes to the case of thresholds, and similar situations where we apply different kinds of moral schemas in different situations, I think there are some real advantages to being explicit, as my theory is, that we actually are using different sorts of first order moral theories to guide our decision making in different circumstances. Being this explicit allows us to make much more sense of several moral judgements we place great faith in that also seem very contradictory with each other. It allows us to reconcile and justify these conflicting intuitions without demanding that we give them up.

The problem of incommensurables

To illustrate this, I will introduce two more common situations where I believe that, much as in catastrophic cases, consequentialist reasoning rather than rule-based reasoning is the right sort of first order theory to apply. These cases are when considering large scale policy and cases of low risk. I will then discuss the intuition most of have that certain of our values are incommensurable with each other, an intuition that many think produces a large problem for consequentialists, and show how SOC can explain and justify that intuition.

In addition to catastrophic cases, there are also other scenarios in which the considerations indirect consequentialists use to argue in favor of common sense moral constraints do not apply, or at least are less weighty. One such example are cases where we are making decisions that affect a large number of people over a long period of time, where the results of our choice are not immediately apparent. The most standard example of this is when deciding what sorts of government policies we should support or advocate for. In this case, in order to make a meaningful decision of any kind it is necessary to weigh a lot of data and take our time to analyze all the potential competing factors, in addition, the consequences generally play out over a long period of time, so we already must take long-term consequences into account. Some of the considerations in favor of abiding by constraints, such as the danger of unintended consequences, still apply, but overall the case for such constraints is weaker.

Another kind of case where we both might have reason to and I will argue in fact *do* evaluate with different moral standards than those we use for ordinary cases are cases where there is a *low risk* of something bad happening when taking a certain course of action. Despite being very different on the face of it, these cases of small-risk are in fact more similar to what we have so far referred to as ‘catastrophic’ cases than they are to ordinary cases. Many of the

arguments that indirect consequentialists have for stringently following rules rather than weighing consequences do not apply to these cases, though often for the opposite reason than in catastrophic cases – that is, because the consequences are less dire rather than because the decision costs are relatively lower. There is also a sunk-cost effect to consider: it plausibly leads to better long-term consequences to adopt a rule such as 'do not commit murder' but it is simply not *practical* to adopt a rule such as 'do not take any action that would risk harming another' because that would preclude so many actions as to make it an impossible rule to internalize. It is also the case that only when large numbers of people are involved do small risks translate into significant consequences, so when evaluating what sorts of general policies we should take towards cases of small risk the decision cost is indeed relatively low, as with other policy decisions. Further, it can be argued that we not only have the intuition that we should set aside absolute moral norms in cases involving extremely large numbers of lives, but we are also on a regular basis willing to *risk* the death of someone to save sufficiently more people using primarily consequentialist rather than deontological reasoning. That is, as a matter of fact our intuitive approach towards cases of small-risk is more consequentialist than deontologist.

My argument here owes much to Alistair Norcross (1998), who notes that we are often willing to risk both our lives and those of other people for various benefits, and that this seems at odds with *absolute* prohibitions on murder for the sake of lesser benefits (let alone to save lives). To give my own example along these lines, imagine the case of the ambulance driver. The ambulance driver knows that there is a person in an accident who will surely die unless they are picked up and returned to the hospital where they will be saved, but also knows that in the process of driving to the location of the accident and back the driver is incurring a (let us suppose) 1-in-a-million chance of killing a pedestrian. Should they incur that risk in order to save

the person in the accident? Most of us would think they should – or rather, that is precisely the decision that ambulance drivers regularly face in real life, but no one is in favor of banning ambulances and almost everyone thinks that their existence is a good thing.

One might respond that to risk a large number of deaths is not as bad as to kill (and that therefore to risk a small chance of death is not one millionth as bad as to take a life), but again our real-life attitudes do not bear this claim out. We regard someone who deliberately engineers a scenario where their victim has a 90% chance of dying as no less a murderer for only creating a risk and not a certainty, and a bioterrorist who infects a population of thousands with a disease that kills one in a thousand is no less a mass-murderer for the fact that the slim possibility exists that no one dies. Further, I think our intuitions hold that introducing such a disease into a population of several thousand is worse than killing one person (though of course both are heinous acts to be avoided if at all possible). Sufficient negligence amounts to negligent homicide and producing large-scale risks for no good purpose is surely something to be avoided. To further the argument along lines similar to Norcross, all of this suggests that we *do* think that large enough risks of death amount to killing, and that enough risks of death amount to a life lost. Thus, while risking a 1-in-a-million chance of killing someone to save 1 person is not necessarily *precisely* the same as killing 1 person to save a million, they do seem to be actions justified by the same consequentialist principles, rather than by deontological norms. That is, these actions are permissible because the benefits far outweigh the costs, but it is otherwise true that we should strive to reduce risks of harm or death to others as much as strive to avoid harming or killing them. Though our duties in this regard are clearly less stringent, the moral factors involved are also proportionately less weighty so that is not a surprise.

Norcross uses arguments like these to argue that principles like "do not sacrifice a life for minor harms or benefits" should be entirely abandoned, but I do not agree with him on this point. Like Norcross, I think these facts – that we allow consequences to outweigh absolute prohibitions when they are large enough, and that we are willing to disobey these sorts of restrictions in cases of very low risk – show that our intuitions *do* allow for consequentialist considerations to override absolute moral rules even in common situations and not just in cases of so-called moral 'catastrophes'. That is, I think the example of the ambulance driver shows that we do not merely override those rules in rare, catastrophic circumstances (as someone like Nozick might say) but also in common, everyday circumstances (it's hard to get any more everyday than driving to work). Unlike Norcross, however, I do not think it therefore follows that we should disregard these rules or cast them aside. All the aforementioned arguments for abiding by moral rules in *small-scale but high-risk* cases still apply: decision costs, the dangers of underestimating adverse consequences, sticking to well-worn rules when faced with situations of incomplete information, and so on. Thus, I think we should keep principles such as "do not kill people to save more people" and "no amount of minor inconveniences saved is worth the loss of a single life" as general rules to abide by – I only advise that we should be aware that we do allow these principles to be overridden when the benefits of doing so are very much larger than the costs, and that this is not a bad thing or a sign of inconsistency but a sign that it is appropriate and justified to adopt different moral frameworks in different moral circumstances.

Another argument along similar lines is made by Tom Dougherty, who argues that non-consequentialism cannot properly accommodate our judgements about risk imposition (Dougherty, 2013). Dougherty highlights two common-sense principles which he argues are inconsistent:

i. *Many-Few:* You may not provide small benefits to many people rather than save the life of someone else. (All else being equal.)

ii. *Risk Tolerance.* You may expose someone to a negligible risk of death in order to otherwise provide this person with a small benefit. (All else being equal.)

The first is of course a close cousin of Norcross' *Worst*. Dougherty argues for their inconsistency as follows: imagine a person dying of poison and you have two medicines that might save them, one of which will certainly do so and the other of which will *almost* certainly do so (with only a one-in-a-billion chance of failing) and will also cure their poison-induced headache into the bargain. By (ii) it seems that we are permitted to give them the second medicine instead of the first. But now if there were a billion people dying of the poison, when treating each individual person it would seem that we are still permitted to give them the second medicine, even though it would be extremely likely that one or more of those billion people would die. It seems therefore that we have rejected (i). Dougherty claims that what makes his argument differ from Norcross' is that it does not rely on consequentialist priors such as the transitivity of 'better than', and therefore that nonconsequentialists have less grounds to reject it.

Dougherty considers and rejects some possible rejoinders, the most notable being that what is permissible to do occasionally is not permissible if it will affect a large number of people, see e.g. (Parfit, Reasons and Persons, 1987, p. 75). This, he says, is to misunderstand the nature of risks, which should be calculated independently of each other: because the trade-off of each instance of risk-taking is independent of every other instance of risk-taking, if the sum of the conveniences is not worth the sum of the risks, then each convenience must not be worth each risk. To think otherwise is to commit the Gambler's Fallacy, thinking that if you get heads nine time in a row the next time *must* be tails: risks do not actually combine like that.

James Kirkpatrick (2018) defends the consistency of holding both *Many-Few* and *Risk Tolerance*, arguing that Dougherty's argument relies on trading in the ambiguity of permissibility, which is used interchangeably on both the narrow and wide scope in Dougherty's argument. But permissibility, says Kirkpatrick, does not agglomerate: a course of action that combines individual permissible actions need not be permissible. By making this distinction, Kirkpatrick is able to avoid inconsistency, but where I think his argument does not provide a very satisfactory answer to the problem is that he does not provide a good enough reason for us to *make* that distinction in this case. In answer to Dougherty's point about the independence of risks, Kirkpatrick is willing to say that it is alright for a motorist to go without a helmet 'just this once' but that they should not make a habit of it. The issue I had with that attitude, though, was not that it leads to a logical contradiction – I am willing to accept that Kirkpatrick is right that it doesn't if we are careful about the scope of permissibility. My issue is simply that it is not an attitude towards wearing helmets that I am willing to condone.

What Kirkpatrick shows is that if we distinguish between the narrow and wide scopes of permissibility we can avoid the logical problem Dougherty poses. But this is to solve only half of the problem, and to my mind the less interesting half. It is one thing to show that holding both (i) and (ii) *need not* lead to a logical contradiction, it is another thing to give a principled reason for us to say that permissibility *should not* agglomerate when it comes to risks. Providing such a reason is, in my view, the real challenge that Norcross/Dougherty arguments produce for nonconsequentialists, and this is where Kirkpatrick's response fails or at least is incomplete. Such reasons are, of course, the province of second order theories, and I think my theory does give us an explanation for why we have intuitions such as (i) despite often being willing to trade risks for convenience. A second order nonconsequentialist has the opposite challenge and it

seems to me a steeper hill to climb, for the same reason as we discussed in the last chapter: they are faced with a problem of deep arbitrariness caused by incommensurable values, a problem the consequentialist does not have. That said there have been various attempts to bridge the gap, with a contractualist approach for example (Dougherty, 2013, pp. 14-16), but I won't go over them in detail here, save to note that such a theory is what is really needed to truly answer the challenge of our conflicting intuitions about risk.

Once we admit the possibility of there being multiple first order moral theories – which are, remember, the sorts of moral theories that we use to guide our decision making – that are each applicable in different circumstances, many of our contradictory intuitions begin to make sense. The contradiction Norcross highlights is one such: the principle that no amount of convenience is worth the loss of human life is one we seem to routinely ignore every time we drive a car. However, such a principle is an invaluable guard against making decisions that appear to be justified in the short term but which are not in the long term. For this principle to have its full effect in preventing adverse consequences, we must do more than pay lip-service to it: we must internalize it, as we have been discussing. And most of us, in fact, *have* internalized it, and thus react with horror to the trading of human life for profit even if, in a sense, we do the very same thing on a regular basis when we introduce minor risks to ourselves and others for small gains. The error here is not in the rule itself, but in applying the rule to circumstances where a different first order theory holds sway.

I do not want to excuse or explain away *all* of our contradictory intuitions here. Some of the reason we seem to be inconsistent about the principles behind our decision making has to be because our intuitions are themselves muddled. We are simply not very good at assessing or incorporating risks, or of properly managing and incorporating all probable outcomes when we

make decisions. Similarly, we don't treat large groups of strangers the same way we do small numbers of people we can see, as we discussed at length in the Demandingness chapters. Just as with our intuitions about demandingness, I don't think all our intuitions about when to apply the rule that human life is incommensurable with other gains are correct; indeed, I think many are simply wrong and should be discarded. And just as with demandingness, I think the great strength of a second order theory is that it gives us the means by which we can evaluate our intuitions in a principled manner. My theory is, once again, revisionary but less so than it might at first seem. I think we *do* utilize different moral rules in different moral circumstances already, but I also think we do so in a somewhat haphazard manner and that a good second order theory will go a long way towards clarifying what sorts of moral theories should be applied to which circumstances. That clarification need not necessarily lead to precise boundaries – as already discussed there are good second order consequentialist reasons to keep some thresholds vague – but it should prevent us from misapplying the wrong sorts of rules in many of the cases we face on a regular basis.

In summary, that certain goods are incomparable with other goods, and in particular that the value of human life is not commensurable with any amount of convenience, is a very good principle to internalize into our decision making for many situations but does not apply to all circumstances. Crucially, this need not lead to contradictions so long as we are careful to recognize that different sets of rules can apply in different circumstances. One very obvious case of differing circumstances is the ordinary vs catastrophic distinction that deontologists have long been concerned with. But I also think that we ought to (and, less importantly, do) apply different and lesser constraints with regard to actions that create small risks of harm compared to actions that create large risks, even if the ratio of harm to benefit remains constant. Finally, one very

interesting case where I think we are and should be at least more straightforwardly consequentialist in our decision making is when making decisions regarding large scale policies. This idea is the subject of the next section.

Policy consequentialism and 'Government House Utilitarianism'

One area that often combines both large numbers and small risks is deciding government policy. Healthcare policy, environmental policy, and many other cases such as deciding speed limits are cases where we are trading small risks distributed throughout the population for small benefits similarly distributed. If I am right in that our ordinary common sense morality, which generally has harsh constraints against trading human life for benefits or even for other lives, does not apply to cases which involve large numbers of people or small risks, then it follows that it also does not apply to cases of policy decisions such as the above. I am willing to say that we ought to be more consequentialist when it comes to policy decisions, or at least that certain constraints that apply to ordinary decision making such as the incommensurability of human life with other goods do not apply in that situation. But I want to also make it clear what I am *not* saying by this.

The idea that consequentialism (or utilitarianism) is as a decision-making procedure (as opposed to a justificatory theory) more apt for deciding public policy than for everyday decision making is one that has an old history, dating back to classical utilitarians such as Bentham (1789) and Mill (1861) but also more explicitly with Sidgwick (1874). The form that this idea took, however, was not always the most appealing. The way it is often put is that classical utilitarianism is 'self-effacing', in that these philosophers seem to be saying that the utilitarian should lie about being a utilitarian and convince others to not be utilitarians, while holding onto utilitarian principles secretly without letting most people know of them. Sidgwick writes "a

Utilitarian may reasonably desire, on Utilitarian principles, that some of his conclusions should be rejected by mankind generally; or even that the vulgar should keep aloof from his system as a whole, in so far as the inevitable indefiniteness and complexity of its calculations render it likely to lead to bad results in their hands."

This attitude towards applying different moral principles to different situations strikes many as being both self-contradictory and extremely elitist: Bernard Williams famously refers to it derisively as 'Government House Utilitarianism' (1985) and characterizes it as arising from the colonialist and imperialist attitudes of the time Sidgwick was writing in. I think there is a lot of truth to that criticism, and I am not comfortable with the idea that the 'true' moral theory is something to be carefully kept from the general public lest they misuse it. There are interesting questions, to be sure, about how much (if at all) we should defer to ethical experts, what their proper role is, and if there is indeed any such thing as ethical expertise at all (Dienhart, 1995) (Yoder, 1998). These are questions I am largely going to sidestep, however, as what I mean when I say that there might be good reasons to adopt different moral principles in different circumstances is definitely *not* that some of us should be using different moral principles than others. Sidgwick's argument that utilitarian calculations done badly likely lead to bad results is something that applies to all of us, not merely the 'vulgar', and is the main argument in favor of indirect or rule consequentialism in most situations.

My view is thus closer to that of R.M. Hare (1981), who argues that there are multiple levels of moral thinking, an idea which should be familiar to the reader by now. Despite the rather unfortunate nomenclature – Hare refers to critical thinking done by 'archangels' and intuitive decision making done by 'proles' – these are not meant to be done by two different

groups of people but by the same person in different circumstances. My critical difference from Hare is that I think he makes two mistakes.

The first of these is not developing a proper distinction between orders. Hare envisions intuitive thinking as us acting according to principles recommended to us by our critical thinking when we don't have the time or information to apply critical thinking directly, which is familiar indirect consequentialism, but he also envisions the critical and intuitive modes of thought as happening simultaneously. This leads to objections by Williams (1988) and Mackie (1985) that criticize the theory on the basis that you cannot simultaneously internalize what Hare calls the intuitive mode of thinking and *also* the critical underpinning. As Williams says, "you cannot combine seeing the situation in that way, *from* the point of view of those dispositions, with seeing it in the archangel's way, in which all that is important is maximum preference satisfaction, and the dispositions themselves are merely a means to that." (Williams, 1988, p. 190). Hare would be better served with a more robust notion of internalization such as that which rule consequentialists like Hooker have, but he resists such a notion (Price, 2019); the reason he does, I think, comes down to his second mistake.

What Hare has in mind with his levels of moral thinking does seem to be the difference between the moral theories we use the make decisions and the underlying justification for those theories – what I have been calling the distinction between first and second order theories – but he clearly *also* thinks that there are cases where we apply the critical 'archangel' mode of thinking to decision making directly. In other words, Hare's theory is very similar to mine, but he equivocates between whether the 'critical' mode is *second-order* utilitarianism being applied to decide what sorts of moral principles to internalize and deploy when we are in the 'intuitive' mode, or *first-order* utilitarianism being applied to our decisions directly when there is

information and time for us to make decisions in that way. He clearly thinks both, and is not wrong to think both, but errs in *equating* the two. This is what opens him up to Williams' criticism.

Let me sum up my position, which will hopefully make the point clear. I am a second order consequentialist, which is to say that I think we should adopt and internalize the decision making procedures that lead to the best consequences in the long run when adopted. I am convinced that for many ordinary situations, the best long term consequences come from internalizing quite nonconsequentialist rules rather than being consequentialist on the first order level, for reasons discussed extensively by now. However, I also believe that there are other situations where we *should* be consequentialists at the first order level as well, and this too is justified by second order consequentialism. What I do not believe, in answer to the Williams-style worry, is that there are situations where we should internalize and deploy two (or more) incompatible sets of moral dispositions: the second order theory that tells us to apply different first order theories for different circumstances (if, as I believe, it tells us that) also tells us which circumstances to apply them to, separately.

Nor is this government house consequentialism: it is true that government policy is something that I consider to be a paradigm example of circumstances where the correct first order theory to apply is more consequentialist, but that does not mean that it is the business of the government house alone. Remember, the entire argument of Chapter 10 was that we *all* have an obligation to be part of the most beneficial group action that we can join: that means advocating for and supporting policy positions, it means organizing, campaigning, protesting, voting. And when we make the decisions about what policies to support or oppose – decisions that we all have to make, not just some of us – it is my belief that we ought to be more

consequentialist than we are in most ordinary situations. Specifically, I believe that at least *some* of the rules we internalize for ordinary cases are wrongfully applied in cases of government policy. One example is the one Norcross argues against, that no amount of minor harms added together can ever equal a human life. And the reasons I think that are second order consequentialist ones: I think that this principle, when applied in the case of government policy, leads too often to more and unnecessary suffering. But what it does *not* mean is that the elite of society who are in positions of power have the privilege to disregard 'vulgar' ethics that the common folk must abide by, an idea that I agree with Williams is abhorrent.

In fact, I would go a bit further than that: when I say that the correct moral theory to apply to questions of government policy is straightforward act consequentialism or at the very least is more consequentialist than the theory we use in many everyday circumstances, that theory is a theory of *ideal* deliberation. It is the theory we use for deciding which set of policies are the ones we should endorse. But when it comes to the practical exercise of policy making in the actual world, there are good reasons for there to be strong limits and constraints on the exercise of state power. I made in the earlier chapters the analogy between rule consequentialism and legal rights such as freedom of speech, and I do believe in strong, enshrined constitutional rights. That is different from trying to apply principles like *Many-Few* to deciding government policy, which is what I believe to be misguided.

Conclusion

In the last chapter, we looked at the case of thresholds: that most people have the belief that deontological constraints are relaxed when the consequences are sufficiently dire. I argued that SOC is better positioned to explain and clarify our intuitions regarding thresholds than threshold deontology unsupported by any second order theory. In this chapter I argue the same

about our intuitions regarding cases where there are low risks of harm but we are dealing with large groups of people (such that there is a high chance that *some* harm will be done). Once again we have very conflicting and contradictory intuitions regarding these cases, and I again think that firstly some sort of second order theory is necessary to adequately clarify cases like these and also that SOC is well positioned to do so. Taking both this and the last chapter together, it appears that we may have good reason to think that our second order theory will recommend different first order theories for different circumstances, and that the reason our intuitions seem so inconsistent is that we have different intuitions about different types of circumstances that are each correct in their own sphere but have no good framework for what happens when they clash. Second order consequentialism can provide that framework.

Chapter 13: Relationships, Honoring, and Virtues

Introduction

In the last couple of chapters I have been exploring the possibility that second order consequentialism might justify not one single first order moral theory, but rather multiple different first order theories that are each applicable to different circumstances. I was initially inspired to pursue this possibility by the puzzle of threshold deontology, and then generally expanded it to other kinds of cases where our nonconsequentialist and consequentialist intuitions clash with each other. In this chapter, I want to try and explore another long-standing problem in ethics that I think we can start giving an answer to once we admit the possibility of there being multiple first order theories. This is the problem that most of the moral theories we discuss as ethicists – such as consequentialism, deontology, and virtue ethics – seem to *not* be what guides our actions when we are interacting with other people with whom we have close personal relationships (Stocker, 1976) (Keller, 2007). In fact, we have the strong intuition that acting according to these theories is *wrong* when it comes to these circumstances, or at least that it would mean that we are acting according to the wrong motive. In this chapter I will discuss this problem in more detail and explore the possibility that we may need a new kind of ethical theory to deal with these cases. Using ideas I have developed in previous chapters – including the possibility of multiple first order theories, the concepts of internalization and binding, and the idea of group action and collective responsibility – I will explain how I think the problem should be addressed in a second order consequentialist framework. Finally, I will discuss how this same approach can be used to explain why certain values should be honored rather than promoted, and why this is not as incongruous with a consequentialist worldview as it is sometimes thought to be.

The problem of self-effacing moral theories

In the last chapter we touched on a criticism of classical utilitarianism being 'self-effacing', namely that there seem to be utilitarian reasons to not profess or even to not act according to utilitarian principles. Some utilitarians like Sidgwick are quite explicit about this, which is why classical utilitarianism is most prone to get this criticism. But Michael Stocker argues that *all* modern ethical theories, be they consequentialist or rule-based, suffer from a similar problem (1976). His argument is quite similar to Williams' 'one thought too many' objection discussed back in Chapter 4, and like Williams he also thinks that the problem arises mainly in the context of close human relationships. Stocker argues that ethical theories give us the wrong *motive* when we are interacting with other people. Imagine if you are a sick person in a hospital who was being visited by a friend, and they told you that they were visiting because they believed that it was their duty, or because doing so would increase the general happiness. You would not be pleased by any such response, and would rather hope they were visiting because they were your friend and not for any high-minded ethical reason. Stocker calls this problem the 'schizophrenia' of modern ethical theories.

Because Stocker focuses on utilitarianism and deontology as examples and considers part of the problem to be that these theories do not adequately capture the value of personal relationships for a life of eudemonia (*Ibid.* p. 460), this argument is often taken as being in support of some kind of virtue ethics. However, Simon Keller points out that virtue theories are hardly immune to the charge of schizophrenia (2007): after all, my friend in the hospital is not likely to be any more impressed with me if I tell him that I am visiting him because it is what a virtuous person would do than because it is my duty. One might respond that when virtue ethicists recommend doing what a virtuous person would do, 'what a virtuous person would do'

is to be read *de re* rather than *de dicto* (Williams, 1995). That is, it should be read as 'do it for the reasons the virtuous person does it' rather than 'do it because that is what a virtuous person would do'. Keller, however, notes that similar moves can be made by consequentialists and deontologists. In general, Keller does not think that virtue ethicists cannot address the problem of self-effacement, but rather that they have no special advantage compared to consequentialists or deontologists.

More recent attempts by virtue ethicists to defend against the charge of self-effacement also admit that similar moves can be made by other types of ethical theories. For example, Glen Pettigrove (2011) tries to answer the charge by making a careful distinction between target of the virtue and the criterion of right action. To use the concrete example of the hospital visit: what makes visiting your friend in the hospital the right thing to do is that it exemplifies the virtue of *philia*, of being a good friend. But what it *is* to be a good friend in this instance is to visit your friend in the hospital because you are worried about them or concerned for how well they are doing and so on. This distinction is rather similar in essence to the *de re* vs *de dicto* distinction drawn by Williams, and Pettigrove also agrees with Keller that non-virtue ethical accounts can similarly avoid the charge of self-effacement by making a careful distinction between the different types of ways in which an action is justified. As we discussed in chapter 4, drawing such a distinction was part of the reason I want to draw a sharp distinction between the different orders, so I am also in agreement on this. Despite not being virtue-ethical at all, my solution to the problem of self-effacement is of the same spirit as Pettigrove.

But before we dive into possible solutions to the problem of self-effacement, I think it is worth taking a moment to narrow down on what kinds of relationships we are talking about. The problem – that ethical theories do not seem to give us the right motive for acting on behalf of

others – does not arise with all of our relationships with each other, but only certain kinds. Some of the relationships we enter into with other human beings are transactional in nature, such as those between a debtor and creditor, or boss and employee. With these types of relationships the 'schizophrenia' Stocker talks about doesn't seem to arise. If someone were to ask why I paid someone back the favor I owed them and I answered that I hold as a general principle that one must repay their debts, I don't think that would be a strange response. If my boss asked me why I worked so diligently and I answered that I think by agreeing to a job one then has an obligation to perform it to the best of their ability, that does not seem very strange either. But if I were asked why I was helping out a friend (or visiting them in a hospital, to use Stocker's example) and I answered that I obey the rule that one should help out one's friends whenever they can, that *does* seem at least a little uncomfortable, like I am depersonalizing or even, in Stocker's terms, de*humanizing* my friendship.

Social psychologists call the difference between types of relationships that I have drawn here the distinction between 'exchange' and 'communal' relationships (Clark & Mills, 1979), and I think a lot of the apparent 'schizophrenia' that Stocker speaks of, the reason why we are uncomfortable with people being motivated by ethical theories to aid their friends rather than the friendship itself, is because it makes our relationships of the latter type too much like the former. There's a lot of evidence that people become uncomfortable or unhappy when relationships of one type are treated too much like they are relationships of another type, such as when married couples keep too close a track of the favors they owe to each other (Buunk & Van Yperen, 1991). It is worth noting that this is *not* the same as saying that fairness isn't important to people who are close friends or married, as it still very much is (Rapson & Hatfield, 2011), but these do seem to be different ways we relate to each other.

And of course, there are other types of ways we relate to other people. If I were to see someone drowning and dived in to rescue them, and answered questions as to why I did it by saying that I believe we have a duty to try and rescue people in such situations, that doesn't seem too strange. Nor would it be very strange if I answered why I was donating money to charity by appealing to my belief that doing so would help to maximize the general wellbeing. It seems to me that a large part of the reason my friend in the hospital would be upset with me is precisely the implication that their case is like these other cases: the implication that our relationship is of the wrong *type.* They are not upset because it is *always* illegitimate to interact with other people while motivated by abstract ethical theories. Rather, they are upset because we only treat people we do *not* have close personal relationships with like this, and they are displeased by the implication that we are not really treating them as a *friend*. Stocker also focuses on love relationships, friendships, and so on, what we might call *close personal relationships*.

The reason why I bring up that we can at least sometimes correctly interact with other people while motivated purely by ethical theories is that it brings up again the possibility of *circumstantial* first order theories. I think that the 'schizophrenia' Stocker describes is more due to us applying the wrong sorts of theories to the situation rather than those theories being wrong when applied in appropriate circumstances (such as when dealing with strangers). If true, this suggests that one way around the problem of self-effacement, where a moral theory recommends that you ignore or act contrary to itself, is to use different moral theories for these different situations. In this case, neither would be self-effacing or deceptive, because the moral theory is not recommending you to act against its own principles. Rather, some second order theory correctly prescribes two different first order theories for different circumstances. This of course would replace the problem of self-effacement with the new problem of describing how these two

theories fit together and what their spheres of applicability are, but we've begun developing the framework for just that over the past two chapters. Just as the disconnect between our consequentialist intuitions about large scale cases and our nonconsequentialist intuitions in small scale cases might be due to us rightly applying different sorts of first order moral theories to different cases, so might the same be true of the disconnect between the sorts of reasons for acting that we think are legitimate when in close personal relationships with other people compared to when we are dealing with strangers.

Second order consequentialism and relationships

By this point it is probably not surprising that I think the problem of self-effacement is to be addressed by expanding on the concept of 'internalization' introduced by Hooker that we discussed back in Chapter 4. However, Stocker considers indirection as a strategy for addressing his objection and remains doubtful, so it is worth seeing why he thinks it does not work. Stocker has two main worries about indirection (Stocker, 1976, p. 463): firstly, that 'there is the great risk that we will get the something else, not what we really want' and therefore that a theory of indirection has to explain the relationship between our motives and the real goal. Secondly, and more importantly, indirection is implausible in this case because we do not act by indirection with regards to love or friendship, but rather 'in these cases our motive has to do directly with the loved one, the friend… as does our reason". If we say that there is some motive beyond love for others that is motivating us to act, then we are both cheapening that love and seem to have to engage in a degree of self-deception that brings us right back to the problem of self-effacement.

Some of these self-same worries were a large part of my motivation for moving from indirect to wholesale second order consequentialism. As we discussed, traditional indirect consequentialists like Sidgwick embrace self-effacement and so do not so much avoid this

problem as bite the bullet. I do not think, however, that distinguishing between first order and second order theory necessarily involves *deception*, self or otherwise. Instead, it means drawing a proper distinction between different types of moral theories: different types of justification, as Pettigrove puts it. There are some crucial distinctions to be made here that make it so that I do not think Stocker's second worry is as big as it seems. What the second order theory justifies is not *particular* friendships but the general practice of seeking out and making friends, and when considered that way it is neither deceptive nor self-effacing. If I said that I wanted to meet some new people and make some new friends because like all humans I am a social animal and require meaningful relationships to have a happy life, this is only strange because it makes explicit the implicit: it is merely a long-winded way of saying I am lonely. The impartial version of this is 'our second order theory encourages adopting and promoting the practice of friendship because humans are social animals etc.' and I will later produce an argument for this not being as strange as it may seem either. For now, though, I want to focus on the other half of the distinction: what justifies *particular* friendships.

The reason a person would want to make friends is that humans are creatures that need such relationships, but the reason they would make the friends they *actually make* depends on the particular friends themselves. They might share common interests or hobbies, have a significant history that binds them together, or simply enjoy each other's company. My reasons for making the actual friends I have do not come from my ethical theory, and neither, therefore, do my reasons for acting on their behalf. Again, my view is similar to Pettigrove despite being not at all virtue ethical, as like him I think *any* answer is going to proceed along similar lines. Here is how he puts it:

> Thinking about the criterion of good action, in this case, will involve thinking about Jones [the friend] and his relationship with her. The criterion is in this sense transparent, since grasping the criterion involves seeing through it to the concrete facts of the case… And thoughts about these facts are precisely what will move Smith when he is paying the best sort of visit to Jones. (Pettigrove, 2011, p. 201)

There are two different sorts of goals being conflated here that need to be carefully distinguished. The reason I seek out friendships is to improve my wellbeing and the wellbeing of others, because I think the world is a better one when people can make real friendships; again, I don't think this is that strange. But this does not motivate me to make the specific friends I actually make, only to seek out friendships in the general sense. It might also motivate me to strive towards the ideal of 'being a good friend' in the abstract, for example by inculcating certain habits in myself. But the reasons I am friends with the *actual people* I am friends with (as opposed to some other people who might have done just as well to satisfy the general goal) are those actual people themselves; consequently, these are also the reasons I act for when I am acting *as* their friend. Stocker worries that "once we begin to believe that there is something beyond such activities as love which is necessary to justify them, it is only by something akin to self-deception that we are able to continue them" (Stocker, 1976, p. 463) but what justifies such activities is not something *beyond* love – it is that humans are creatures that need love.

The 'self-deception' worry seems to imply the thought process is something like this: I think the best world (let's use indirect consequentialism as the example, though this general worry appears for all indirections) will come about when I seek to be friends with people; in order to be a real friend (and thus bring about the best world) my actions cannot be motivated by the general good; therefore I need to 'forget' my 'true' goal in order to best bring it about, which

seems to require active and deliberate self-deception. But that is missing a few stages in the process, to my mind. When I became friends with that person – not later, when I am acting on that friendship by visiting them, but when I *entered into* the relationship – I took on that person as a source of reasons for acting. Again, this is similar to the idea of internalizing a rule developed by Hooker, or of binding oneself to a promise, both of which are ideas I discussed back in Chapter 4. From that point on, I no longer think about the second order reasons for which I sought out friendships. But this is not because I am forgetting those reasons or deceiving myself, but because those reasons simply do not apply to the particular actions I take on behalf of my actual friends. Rather, they charge me to seek out friendship and to be a good friend in the general sense.

As to why becoming friends with someone involves internalizing them as a reason for action in this way, I think that this is what a friendship (or any meaningful personal relationship) *is*. It is precisely this that makes such relationships meaningful: coming to care about another person for and as themselves, rather than because of some more general motive. As to the second order justification for why we ought to make such relationships, I think again the answer must be constitutive: humans are social animals. It is precisely these sort of relationships, where we care for each other as individuals and not for obligation's sake, that people *need* (to varying degrees depending on individual difference, a point we'll come back to later) to be healthy and happy. Just as the act of binding oneself to a promise is what gives promise-keeping its value as a practice, committing to care about another person for themselves and not for external reasons is what makes our relationships with them meaningful. Consequentially, it is precisely what our second order theory tells us to do when we enter into close personal relationships with others.

There is no element of deception here because I am doing this with a clear eye: my second order theory motivates me to enter into such relationships in the general, personal and specific reasons motivate me to enter into the particular relationships I form, and what it *is* to enter into a relationship like that is to internalize the other people involved as sources of reasons for action. If there is anything strange about this, it is that it makes explicit what we generally don't think about in nearly as much detail. Again, humans are social animals, and as such we form such relationships for the most part naturally and instinctively. However, when I reflect on the actual process of forming such relationships, I do think this is essentially what is going on. There is no contradiction between my second order motives to make friends and my motives for visiting a friend in the hospital because these are operating at different levels of reasoning – at different parts of the process – and thus no need for any kind of deception. Again, this is the difference between first order and second order moral theories; and if I am right, then the first order theory that governs our actions regarding close personal relationships such as friendship is not first order consequentialism or deontology or even virtue ethics, but rather a kind of person-centered particularist moral theory.

One thing I do agree with Stocker about is that this first order theory requires a lot more explication and exploration (*Ibid.* p. 460). Traditional first order theories have generally regarded special obligations as a kind of *exception* to the general rules we abide by, rather than developing full-blown person-centered ethical theories that govern our close personal relationships. Part of the reason for this is that doing so would seem to involve embracing a thoroughgoing pluralism about ethical theories, at least at the first order level, which theorists are understandably reluctant to do. But as the subject of the last couple of chapters has been precisely that such a pluralism may also be warranted for other reasons, I think such a theory is very much worth exploring. I

will not do so very much here, because the focus of this dissertation is the second order level and the purpose of this chapter more to show how that level relates to our hypothetical first order theory. But hopefully in this chapter I am carving out a space for such a theory, so to speak, which will serve as a jumping off point for others.

Consequentialism about relationships

This brings us back to Stocker's other worry about indirection: the danger that we get what we are aiming for and not what we really want, that the indirect theorist has to explain the connection between our motives and our real goal. At this point, I can no longer talk about second order theories in the abstract (much of what I have said so far can be said by any other second order theorist and is said by Pettigrove) and focus on my theory in the specific. I am a second order consequentialist: I think people should form friendships because I think the world is a better place when it has such relationships in it. Presumably, therefore, this would mean that I think we have a duty to promote friendship in the world. This, however, seems quite a strange thing to say, as we seem to have no such duty. At least, we don't generally think we have a *moral obligation* to go out and make friends and encourage other people to make friends. This seems again to be the wrong sort of motive.

There are two main things I would say to mitigate this worry somewhat: that the best friendships are unforced and that the world is not better for having more friendships. The first point is simple enough: the best world consists of genuine friendships that formed naturally, not by people forming friendships because they believe it is their moral duty to do so. This changes the tenor of what moral obligations we have with respect to friendship: rather than the obligation to make friendships and push others into doing so, what we have is the obligation to *make the*

world such that friendships can form and flourish. This seems to me a much less strange thing to say, as I do think we have such an obligation.

The second point is something that applies to most indirect or instrumental goods: recall that the total utilitarian will still advocate for an equitable distribution of resources, as resources have diminishing returns and so the greatest total wellbeing will come about when resources are equally distributed. Something similar applies here, though there is the additional complication with respect to relationships (which also applies to money and similar but not to the same extent) that different people are quite different. Some people are introverts or asocial, and need fewer meaningful relationships to be happy; others are extremely social and suffer more for lack of human interaction than most. And everyone will reach a point where more friendships are not going to make their lives any better. The best world is not the one that has the *most* friendships in it, but the one in which each person has the number of friendships that are correct for their individual needs. And I do think we have a moral obligation to make that world come about.

The reader may still be tempted to resist this claim, but I think this is to misunderstand the *nature* of that obligation. This is where another concept I developed earlier in this dissertation comes into the play, and that is the idea of collective responsibility and group actions I talked about in Chapter 10. I do not think we have the responsibility to go around and make sure that everyone around us has the correct number of relationships (though we probably should do that too to some extent) but rather that we should try and join into group actions that will bring about the best world. And when we consider what those group actions look like, we see that these obligations are not so strange: they will include things like stamping down on bullying and harassment, making sure we are not throwing up unnecessary societal barriers in the way of friendships, whilst also not placing undue social pressure on those of us who are genuinely

happier alone. It includes being mindful of the media we promote and consume, and how that media depicts relationships. It means, in short, that we have a collective responsibility to shape our society such that people will be able to each individually make the right number of friends for their own personal happiness. Once you see the shape of our responsibilities in this regard, they will not, I think, seem very strange at all.

Honoring and promoting values

This framework for understanding our moral obligations with regard to personal relationships is also one which can be used for many other obligations we have. In this section, I wish to deploy it for a constellation of values which first-order consequentialism traditionally tends to struggle with: values which we best serve by honoring them instead of promoting them. This distinction was first introduced by Philip Pettit (1989) who analyzed it as the difference between consequentialism and nonconsequentialism. Consequentialists, says Pettit, respond to values by promoting them, by attempting to ensure that the world has as much of that value as possible. Nonconsequentialists respond to values by honoring them, so for instance one honors honesty by striving to be as honest as possible within one's own life, rather than trying to ensure that the world has as much honesty in it as possible. Pettit saw this as consequentialists and nonconsequentialists responding in different ways to the same set of values, and that this is thus what made for the difference between them.

But there is another way of viewing this distinction, which is as being between different *kinds* of values rather than different ways of responding *to* value. McNaughton and Rawling (1992) argue that the honoring/promoting distinction is not a good way of understanding the difference between consequentialism and deontology, pointing out that, for instance, this analysis fails when it comes to happiness. The honoring equivalent of the utilitarian's promotion of

happiness would be egoistic hedonism, which most deontologists would not want to adopt as a rule in their systems. Pettit tries to argue that to honor happiness is to make the people around you happy and strive to not actively cause unhappiness, but McNaughton and Rawling point out that this is, structurally, different from the happiness value that utilitarians promote. There is no simple nonconsequentialist alternative to promoting happiness (besides egoism), undermining the claim that the main difference between the two types of ethics is their different ways of relating to the same values. On the other hand, there are certain values that it is odd to talk about promoting, like the value of keeping one's promises: one does not go around ensuring that there are more promises in the world to be kept. It seems to me that it is not that there is one set of values that we relate to in different ways depending on whether we are consequentialists or not. Rather, there are some values we mostly agree are the sorts to promote, some we honor, and some to which different people react differently. The challenge for any ethical theory, regardless of type, is to explain which value should be responded to in which way, and why.

One argument sometimes made in favor of consequentialism, especially maximizing consequentialism, is that the most rational response to value is to promote it. To put it another way, if something is good then it seems to follow all else being equal the more of it the better (Scheffler, 1988, p. 1). You will note, though, that I did not advance this argument when I was arguing for consequentialism back in Chapter 3, and that was for a good reason: I am not sure I believe in it. At most, it would seem to be true only of intrinsic goods, and I am not even sure about that. Many goods have diminishing returns for wellbeing, and even with wellbeing the world with the more wellbeing in it may not be better if it is unequally distributed: that is what it

means to include distribution in one's theory of the good[41]. And even leaving aside diminishing returns, just because the world is better off for having something doesn't necessarily mean the world is better the *more* there is of that something. Sometimes something is valuable in that you need a certain amount of it, but gives no additional benefit if you have more than that amount: oxygen, to give a simple example. Indeed, many of those valuable things can be harmful in excess, including oxygen.

This is where our discussion of personal relationships and our obligations thereof come into play. In the previous section, I explained why I do not think the claim that we have an obligation to improve the world with respect to friendships is actually all that strange. What *is* strange is interpreting that as meaning that we need to promote friendship in the world, and we rightly regard that as the wrong way of relating to the value of friendship. But if one instead interprets it as meaning that we have the obligation to make the world such that people can most easily achieve the number of friendships that they need, and also keep in mind that the responsibility is a collective and not an individual one, all that seems strange about the obligation vanishes.

I think the same is true of most of the values we honor rather than promote: indeed, I think that what *makes* them values we correctly respond to in the former way rather than the latter is their similarities to friendship. Specifically 1) that they are not such that the world is better off the more of the value in question there is 2) that their greatest value is shown when they are unforced and 3) that therefore our obligations with regards to that value is less making it so that there is as much of it in the world as making the world as friendly to it as possible. That

[41] I suppose you could characterize the inclusion of equality in one's theory of the good by saying that all else being equal the world is better the 'more equality' there is, but this strikes me as shorthand for saying you care for how goods are distributed; otherwise it is a strange and oddly recursive way of thinking about 'equality'

last is where I betray my consequentialist leanings, as it could be regarded as 'promoting' the value, in a certain sense. But I think, when we keep in mind the principles of collective actions developed in previous chapters, that our obligations in this regard are quite common-sensical. And not just with regards to friendships, but with many other values. Take art, for example: my obligations with regards to promoting the amount of art in the world are much less about going out and forcing people to paint and much more about voting for certain programs, doing what I can to support artists on a personal level, and trying to help make our society into one that fosters creative expression.

For another example, let us again look at promise-keeping in more detail. I would argue that the reason (or at least one of the reasons, but let us focus on this one for now) that keeping one's promises is valuable is that in engenders trust in a way that makes relationships themselves more valuable. But it is the commitment to keeping one's promises itself that engenders this trust, not the *amount* of promises one has kept – indeed, making too many promises too lightly can cheapen that trust somewhat. In addition, just like we discussed earlier with friendships (and for similar reasons) forced promises do not engender trust in this way, for easy to see reasons. Finally, and again as with friendships, I *do* think we have a responsibility to promote promise-keeping as well as honor our own promises, but the nature of that responsibility is collective and distributed. The responsibility lies on all of us to teach each other about the value of promises and to respect the promises of others, and that is how one makes the world the best place it can be with respect to promise-keeping, rather than maximizing the number of promises kept as such.

Something similar is what I would say for most values towards which the natural response is to honor them rather than promote them. Oddly enough, the more one focuses on values in the specific rather than good in the abstract the less true this claim – that the natural

response to the good is to make as much good as can be – seems to me. Rather, I would say that the consequentialist impulse is to make the *world* the best it can be: but *individual* goods that are such that the world is always better off the more of them there are seem to be in the minority: wellbeing is perhaps the only standout example. Perhaps the difference is between intrinsic and instrumental values, as I am sure the welfarist would claim, but even that might be questioned. Suppose you had some sort of desire theory of value, such that what it is for something to be valuable is that people desire it: well, for most things that we value, it is possible to have 'enough' of it. Even the putative exception, desire-satisfaction itself, seems to have this quality: my instinct, at least, is that it is possible for all my desires to be sated (at least temporarily) at which point satisfying my desires *more* isn't going to make the world better. Now, in realistic scenarios most people have not achieved peak happiness, which is why I think we *should* respond to happiness by promoting it, but this seems now a *contingent* reaction that arises from the state of the world, rather than something about the value itself.

In the end, I agree with McNaughton and Rawling that the honoring/promoting distinction is not the key difference between consequentialism and nonconsequentialism; it rather seems to me a mostly value theoretic question. Our second order theory will tell us, for each value, whether we should promote it or honor it, for whatever reasons the theory gives. In my case, it would depend on whether promoting or honoring the value would best serve making the better world, which in turn would depend on *how* the value makes the world better (directly or indirectly? Is it always better if there is more of it or is it just that there should be enough? Etc.).

Virtue ethics and SOC

A final note about virtues and virtue ethics. Over the course of this chapter, I think I have developed a clear framework for how virtues fit into SOC. Arguably such was developed back in

Chapter 5: virtues, like rules, must be internalized to be effective from a second order consequentialist perspective, and it was in light of the possibility that sometimes the former was better than the latter that I declined to call myself a rule consequentialist. The framework in this chapter answers some lingering worries left behind by the idea of virtues justified by a SOC, such as the worry of self-effacement.

However, it is hardly unusual for a moral theory to feature virtues in this way: just as even most deontologists care about consequences to some extent, so do most ethical theories advocate that we behave virtuously. What distinguishes virtue ethics is the *centrality* of virtues in the picture, not their existence (Hursthouse & Pettigrove, 2018), just as with consequences for consequentialists and rules for deontologists. Do I think that this sort of first order theory, one in which our actions are judged on the *basis* of the virtues demonstrated, is one we should adopt in at least some circumstances? I had, in fact, initially considered it for the theory that governs our relationships with other people. But as we discussed in this chapter I no longer think that virtue ethics has any particular advantage in this regard, and that what we need is some sort of person-centered theory yet to be fully developed. Virtues have a place in my theory, just as with any other ethical theory, but even at the first order level I do not think it is pride of place.

Conclusion

Many first order theories seem to have us acting for the wrong reasons when we act for the sake of those closest to us – our friends, family, and so on. We should not be acting to aid them because doing so would maximize the general good or satisfy the correct moral rule, but rather because they are close to us. Once we introduce the concept of internalization discussed by Hooker, we see that the reason for this so called ‘schizophrenia’ is that most ethical theories do not have us internalizing other people as reasons for action in themselves. Instead, they treat

special obligations as a specific kind of exemption or exception somewhat awkwardly grafted onto their main theory. The reason for this awkwardness is that to commit to an ethical theory that governs our personal relationships separately from our other actions is to commit to first order pluralism. Though this solves the problem of schizophrenia, it is something moral theorists are reluctant to do.

As discussed in previous chapters, however, I already think we have second order reasons to be first order pluralists. Furthermore, SOC provides a unifying framework that underlies the different first order theories, alleviating the main worry of pluralism of determining what to do when the theories clash. Some may resist a consequentialist justification for forming relationships, as they believe that human relationships are the kind of value that is properly honored rather than promoted. However, I think this worry is mitigated once we recognize two things. Firstly, some values do not simply make the world better if there is more of them, but need to be distributed correctly and in the right amounts. Secondly, the responsibility for ensuring that the world is friendly to the value is a collective one. Once we recognize this, the shape of our obligations with regard to relationships and similar values is not strange at all, and SOC allows us to escape from Stocker's charge of schizophrenia.

PART 5: CONCLUSION

At the beginning of this dissertation, I argued that the case for second order consequentialism could not be made by advancing one single decisive example or argument but by applying the approach to several different cases and showing that doing so gives us fruitful avenues of analysis. As a result, my dissertation is not a conclusive argument in favor of second order consequentialism. Rather, it is an argument that we have very good reasons to take the idea of SOC seriously and develop it further. Producing a full-fledged moral theory is beyond the scope of this dissertation and would require at the very least a developed theory of the Good, but I do think that there are some firm conclusions we can make on the basis of the arguments I have defended here:

1) **A second order theory allows us to interrogate and analyze our moral commitments in a systematic fashion**

Many consequentialists are generally skeptical of our intuitions, but as I discussed in Chapter 2 I don't think that is a viable position. However we have many reasons to suspect that our intuitions have particular flaws and systematic biases. A second order theory gives us a framework with which we can evaluate our intuitions. More generally, it gives us a non-arbitrary way of weighing different first order commitments against each other when they clash. As a direct result of both of these things:

2) **We need *some* sort of second order theory.**

This is an especially strong case when it comes to the problems of demandingness and thresholds. In both those areas, we see that without a second order framework to support us we

are left with a deep arbitrariness problem when it comes to gauging the right level of demandingness or the location of the threshold, one that simply cannot be solved on the first order level alone. What is interesting is that the first of these is often seen as a special problem for consequentialism while the second is seen as one for deontology, but I would argue that that is not necessarily true in either case. Demandingness is in fact a general ethical problem, because every plausible ethical theory demands that we sacrifice *something* for the sake of others and thus must answer the question of how much. Similarly, while consequentialists have an easier time dealing with thresholds I think a large part of that is because they are used to thinking in indirect terms. Threshold deontology fails when it operates at solely the first order level. And as I discussed in Chapter 7, answers to the Demandingness objection, such as satisficing consequentialism, fail when they do the same. We need some kind of theory that tells us how much our first order moral theories can ask of us, where and when they apply, and in general the limits and boundaries of those theories. This is especially pressing in the case of demandingness, as we have very good reason to think that our intuitions are especially unreliable there.

3) Second order consequentialism is a particularly promising second order theory

This is because the solutions it provides to the problems of demandingness and thresholds can themselves also be fruitfully applied in other areas. While the theory of blameworthiness it advocates is revisionary, I think it is clear that at least some revision about how we assign blame is indeed due. Further, the theory can also be applied fruitfully to problems of collective responsibility. Similarly, applying SOC to the problem of thresholds opens the door to first order pluralism, which also allows us to address the problem of the wrong motive that arises when most ethical theories are applied to our close personal relationships. And combining that last idea

with the approach to collective responsibility developed earlier gives us a framework for understanding values that ought to be honored rather than promoted in a consequentialist context.

4) **Second order consequentialism allows us to keep many of the benefits of consequentialism while avoiding some of its downsides**

This is especially true of consequentialism's theoretical virtues such as simplicity and unity of explanation. These are particularly desirable theoretical traits at the second order level because one of the benefits of a second order theory is precisely that it gives us a framework to balance the multiple facets of a more complex first order theory. It can even license pluralism at the first order level, as we have discussed in the last few chapters, while allowing us to keep a unifying underlying foundation. But the more complex and multifaceted we make our second order theory, the more we lose some of the main benefits of having such a theory in the first place. This is why I think that SOC is the most promising second order theory, though it is also true that there is a much larger tradition of indirect consequentialism to draw from when compared to nonconsequentialist ethical theories.

Summary

To these virtues of SOC, I would also add that while it is revisionary in many places, I do not think it ever produces a conclusion I find unacceptable. To go back to Enoch's terminology which I introduced in Chapter 1, it loses far fewer plausibility points when compared to first order consequentialism. I think it is reasonable to hold that the virtues of consequentialist theories fail to overcome their flaws at the first order level. But at the second order level I think that they succeed in doing so, to the point that consequentialism is the most promising approach at that level.

There is yet a long distance to go between showing that an approach is viable and promising and actually developing a full-fledged second order ethical theory. But that demonstration of viability and fruitfulness does make me excited to continue with this overall project, confident that such a theory can be developed, and happy to call myself a second order consequentialist.

REFERENCES

Ackeren, M. v., & Sticker, M. (2015). Kant and Moral Demandingness. *Ethical Theory and Moral Practice 18 (1_*, 75-89.

Alexander, L. (2000). Deontology at the Threshold. *San Diego Law Review, Vol. 37*, 893-912.

Alexander, L. (2018). Conclusion: Appreciation and Responses. In H. Hurd (Ed.), *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander* (pp. 407-440). Cambridge: Cambridge University Press.

Alexander, L., & Moore, M. (2020, April 17). *Deontological Ethics.* Retrieved from The Stanford Encyclopedia of Philosophy (Winter 2016 Edition): https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/

Anscombe, G. (1956). Mr. Truman's Degree. In *The Collected Philosophical Papers of G. E. M. Anscombe, vol. III, Ethics, Religion and Politics (1981)* (pp. 62-71). Blackwell: Oxford.

Anscombe, G. (1958). Modern Moral Philosophy. *Philosophy, 33(124)*, 1–19.

Arneson, R. (2018). Deontology's Travails. In H. Hurd (ED.), *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander* (pp. 350-370). Cambridge: Cambridge University Press.

Ashford, E. (2000). Utilitarianism, Integrity, and Partiality. *The Journal of Philosophy, Vol. 97, No. 8*, 421-439.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation.* Garden City: Doubleday. republished in 1961.

Bentham, J. (1840). *Theory of Legislation.* Bristol: Thoemmes Continuum rept. 2004 Translated from the French of Etienne Dumont by Richard Hildreth.

Berkey, B. (2014). Climate Change, Moral Intuitions, and Moral Demandingness. *Philosophy and Public Issues - Filosofia E Questioni Pubbliche 4 (2)*, 157-189.

Bradley, B. (2006). Against Satisficing Consequentialism. *Utilitas, 18*, 97-108.

Bradley, B. (2009). Saving People and Flipping Coins. *Journal of Ethics and Social Philosophy Vol. 3, No. 1*, 1-13.

Brandt, R. (1992). *Morality, Utilitarianism, and Rights.* Cambridge University Press.

Broome, J. (1991). *Weighing Goods.* Oxford: Basil Blackwell.

Buunk, B. P., & Van Yperen, N. W. (1991). Referential Comparisons, Relational Comparisons, and Exchange Orientation: Their Relation to Marital Satisfaction . *Personality and Social Psychology Bulletin, 17(6)*, 709–717.

Case, S. (2016). Rethinking Demandingness: Why Satisficing Consequentialism and Scalar Consequentialism are not Less Demanding than Maximizing Consequentialism. *Journal of Ethics and Social Philosophy 10 (1)*, 1-8.

Chang, R. (1997). Introduction. In R. Chang (ed.), *Incommensurability, Incomparability, and Practical Reason.* Cambridge: Harvard University Press.

Chappell, R. Y. (2019). Willpower Satisficing. *Noûs 53 (2)*, 251-265.

Choen, Y. (2014). Don't Count on Taurek: Vindicating the Case for the Numbers Counting. *Res Publica 20 (3)*, 245-261.

Clark, M. S., & Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology, 37(1)*, 12-24.

Cooper, D. (1968). Collective Responsibility. *Philosophy, 43*, 258-268.

Cosmides, L., & Tooby, J. (1992). Cognitive Adaptions for Social Exchange. In J. Barkow, L. Cosmides, & J. Tooby, *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163-228). New York: Oxford University Press.

Cullity, G. (1994). International Aid and the Scope of Kindness. *Ethics, vol. 105, no. 1*, 99–127.

Cullity, G. (2004). *The Moral Demands of Affluence.* Oxford: Oxford University Press.

Cushman, F., Young, L., & Greene, J. (2010). Multi-System moral psychology. In J. Doris (Ed.), *The moral psychology handbook* (pp. 47-71). New York: Oxford university Press.

Dienhart, J. (1995). Rationality, Ethical Codes, and an Egalitarian Justification of Ethical Expertise. *Business Ethics Quarterly , Vol. 5, No. 3*, 419-450.

Dougherty, T. (2013). Aggregation, Beneficence and Chance. *Journal of Ethics and Social Philosophy 7*, 1-19.

Ellis, A. (1992). Deontology, Incommensurability and the Arbitrary. *Philosophy and Phenomenological Research*, 855-875, Vol. 52, No. 4 .

Enoch, D. (2011). *Taking Morality Seriously: A Defense of Robust Realism.* New York: Oxford University Press.

Ernst, Z. (2007). The Liberationists' attack on Moral Intuitions. *America Philosophical Quarterly Volume 44, Number 2*, 129-142.

Feldman, F. (1997). *Utilitarianism, Hedonism, and Desert.* New York: Cambridge University Press.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: An Essay on Moral Responsibility.* Cambridge: Cambridge University Press.

Foot, P. (1978). The Problem of Abortion and the Doctrine of the Double Effect. In *Virtues and Vices and Other Essays* (pp. 19-33). Berkeley, CA: University of California Press.

French, P., & Wettstein, H. (2014). *Midwest Studies in Philosophy (Volume XXXVIII: Forward Looking Collective Responsibility).* Minneapolis: University of Minnesota Press.

Goodin, R. E. (2009). Demandingness as a Virtue. *Journal of Ethics 13 (1)*, 1-13.

Greene, J., Cushman, F., Stewart, L., Lowenburg, K., Nystrom, L., & Cohen, J. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition 111*, 364-371.

Greene, J., Morelli, S., Lowenberg, S., Nystrom, L., & Cohen, J. (2008). Cognitive load selectively. *Cognition, 107*, 1144-1154.

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion.* New York: Pantheon Books.

Hare, R. (1981). *Moral Thinking: Its Levels, Method and Point.* Oxford: Clarendon Press.

Hare, R. (1981). *Moral Thinking: Its Levels, Method and Point.* Oxford: Clarendon Press.

Harris, S. E. (2015). Demandingness, Well-Being and the Bodhisattva Path. *Sophia 54 (2)*, 201-216.

Hieronymi. (2004). The Force and Fairness of Blame. *Philosophical Perspectives*, 18(1): 115–148.

Hieronymi, P. (2001). Articulating an Uncompromising Forgiveness. *Philosophy and Phenomenological Research*, 62: 529–555.

Hooker, B. (2002). *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality.* Oxford: Oxford university Press.

Hursthouse, R., & Pettigrove, G. (2018, December 1). *Virtue Ethics.* Retrieved from The Stanford Encyclopedia of Philosophy (Winter 2018 Edition): https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/

Jefferson, A. (2019). Instrumentalism About Moral Responsibility Revisited. *The Philosophical Quarterly Vol. 69, No. 276*, 555-573.

Kagan, S. (1984). Does Consequentialism Demand too Much? Recent Work on the Limits of Obligation. *Philosophy and Public Affairs, Vol 13, No. 3*, 239-254.

Kagan, S. (1989). *The Limits of Morality.* Oxford: Clarendon Press.

Kagan, S. (1998). *Normative Ethics.* Boulder: Westview Press.

Kagan, S. (2002). Kantianism for Consequentialists. In I. Kant, & A. Wood (ed.), *Groundwork for the Metaphysics of Morals* (pp. 111-156). Yale.

Kagan, S. (2011). Do I Make a Difference? *Philosophy & Public Affairs, Vol. 39, No. 2*, 105-141.

Keller, S. (2007). Virtue Ethics is Self-Effacing. *Australasian Journal of Philosophy, 85 (2)*, 221-32.

Kirkpatrick, J. R. (2018). Permissibility and the Aggregation of Risks. *Utilitas Vol. 30 Issue 1*, 107-119.

Korsgaard, C. (1989). Personal Identity and the Unity of Agency: A Kantian Response to Parfit. *Philosophy and Public Affairs 18, no. 2*, 101-132.

Lawlor, R. (2006). Taurek, numbers, and probabilities. *Ethical Theory and Moral Practice 9*, 149-166.

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford: Oxford University Press.

Mackie, J. (1985). *Persons and Values: Selected Papers Volume II.* Oxford: Oxford University Press.

Macnamara, C. (2011). Holding Others Responsible. *Philosophical Studies, 152*, 81-102.

Macnamara, C. (2015). Reactive Attitudes as Communicative Entities. *Philosophy and Phenomenological Research, 90*, 546-569.

McElwee, B. (2010). The rights and wrongs of consequentialism. *Philosophical Studies, 151(3)*, 393-412.

McElwee, B. (2011). Impartial Reasons, Moral Demands. *Ethical Theory & Moral Practice, 14(4)*, 457-466.

McElwee, B. (2017). Demandingness objections in ethics. *The Philosophical Quarterly, 67(266)*, 84-105.

McGeer, V. (2014). P.F.Strawson's Consequentialism. In N. Tognazzini, & D. Shoemaker (eds.), *Oxford Studies in Agency and Responsibility: 'Freedom and Resentment' at 50* (pp. 64-92). Oxford: Oxford University Press.

McGeer, V. (2015). Building a Better Theory of Responsibility. *Philosophical Studies, 172*, 2635-49.

McKenna, M. (2012). *Conversation and Responsibility.* New York: Oxford University Press.

McKenna, M. (2013). Directed Blame and Conversation. In D. J. Coates, & N. A. (eds.), *Blame: Its Nature and Norms* (pp. 119-140). New York: Oxford University Press.

McNaughton, D., & Rawling, P. (1992). Honoring and Promoting Values. *Ethics 102, No. 4*, 835-843.

Mendola, J. (2006). *Goodness and Justice.* New York: Cambridge University Press.

Meyers, C. (2014). Brains, trolleys, and intuitions: Defending deontology from the Greene/Singer argument. *Philosophical Psychology, vol 28*, 466-486.

Mill, J. (1861). *Utilitarianism.* New York: Oxford University Press, 1998.

Miller, D. (2007). *National Responsibility and Global Justice.* Oxford: Oxford University Press.

Moore, G. (1903). *Principia Ethica.* Cambridge: Cambridge University Press.

Moore, M. (1997). *Placing Blame: A Theory of the Criminal Law: Oxford University Press.* Oxford: Oxford University Press. Retrieved from Oxford Scholarship Online: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199599493.001.0001/acprof-9780199599493.

Moore, M. (2018). The Rationality of Threshold Deontology. In H. Hurd (Ed.), *Moral Puzzles and Legal Perplexities: Essays on the Influence of Larry Alexander* (pp. 371-387). Cambridge: Cambridge University Press.

Morris, R. (2017). Praise, blame, and demandingness. *Philosophical Studies*, 1857-1869.

Moss, J. J. (2011). Strategies for Defusing the Demandingness Objection. *Philosophy Dissertations, Theses, & Student Research. 6. University of Nebraska-Lincoln*, http://digitalcommons.unl.edu/philosophydiss/6.

Mulgan, T. (2001). *The Demands of Consequentialism.* Oxford: Oxford University Press.

Murphy, L. (2000). *Moral Demands in Nonideal Theory.* New York: Oxford University Press.

Nefsky, J. (2021). Climate Change and Individual Obligations: A Dilemma for the Expected Utility Approach, and the Need for an Imperfect View. *Philosophy and Climate Change*, 201-221.

Norcross, A. (1998). Great harms from small benefits grow; how death can be outweighed by headaches. *Analysis 58.2* , 152-158.

Norcross, A. (2006). The Scalar Approach to Utilitarianism. In H. West (ed.), *The Blackwell Guide to Mill's Utilitarianism.* (pp. 217-32). Wiley-Blackwell.

Nozick, R. (1974). *Anarchy, State and Utopia.* New York: Basic Books.

Nozick, R. (1974). *Anarchy, State and Utopia.* New York: Basic Books.

Parfit, D. (1987). *Reasons and Persons.* Oxford: Clarendon Press.

Parfit, D. (2011). *On What Matters: Volume 1.* Oxford University Press.

Pettigrove, G. (2011). Is Virtue Ethics Self-Effacing? *Journal of Ethics, 15 (3)*, 191-207.

Pettit, P. (1989). Consequentialism and Respect for Persons. *Ethics 100*, 116-26.

Pettit, P., & Smith, M. (2000). Global Consequentialism. In B. M. Hooker, & D. E. Miller (eds.), *Morality, Rules, and Consequences* (pp. 221-233). Edinburgh: Edinburgh University Press.

Podgorski, A. (2018). Wouldn't it be Nice? Moral Rules and Distant Worlds. *Noûs 52(2)*, 279-294.

Portmore, D. (2008). Dual-Ranking Act-Consequentialism. *Philosophical Studies Vol. 138, No.3*, 409-427.

Price, A. (2019). *Richard Mervyn Hare.* Retrieved from The Stanford Encyclopedia of Philosophy (Summer 2019 Edition), Edward N. Zalta (ed.): https://plato.stanford.edu/archives/sum2019/entries/hare/

R.M., H. (1981). *Moral Thinking.* Oxford: Clarendon Press.

Railton, P. (1984). Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs, 13*, 134-171.

Rapson, R. L., & Hatfield, E. (2011). Equity Theory in Close Relationships. In P. A. Van Lange, A. W. Kruglanski, & E. Higgins (eds.), *Handbook of Theories of Social Psychology: Volume 2* (pp. 200-217). London: Glyph International.

Rawls, J. (1971). *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Reid, T., & Brooks (ed.), D. R. (1764/1997). *An Inquiry into the Human Mind on the Principles of Common Sense.* University Park: Pennsylvania State University Press.

Richardson, H. S. (1990). Specifying Norms as a Way to Resolve Concrete Ethical Problems. *Philosophy and Public Affairs, 19(4)*, 279-310.

Rosati, C. S. (2020, March 2). *Moral Motivation.* Retrieved from The Stanford Encyclopedia of Philosophy (Winter 2016 Edition): https://plato.stanford.edu/archives/win2016/entries/moral-motivation/

Sanders, J. T. (1988). Why the numbers should sometimes count. *Philosophy and Public Affairs 17*, 3-14.

Scanlon, T. (1998). *What We Owe To Each Other.* Cambridge, MA: Harvard University Press.

Scanlon, T. (2008). *Moral Dimensions.* Cambridge, MA: Harvard University Press.

Scanlon, T. (2013). Interpreting Blame. In D. J. Coates, & N. A. Tognazzini (eds.), *Blame: Its Nature and Norms* (pp. 84-99). New York: Oxford University Press.

Scheffler, S. (1982). *The Rejection of Consequentialism.* Oxford: Clarendon Press.

Scheffler, S. (1988). *Consequentialism and Its Critics.* Oxford: Oxford University Press.

Scheffler, S. (1992). Prerogatives without Restrictions. *Philosophical Perspectives 6*, 377-397.

Sen, A. (1982). Rights and Agency. *Philosophy and Public Affairs, 11(1)*, 3-39.

Shaw, W. (2000). Between Act and Rule: The Consequentialism of G.E. Moore. In B. Hooker, E. Mason, & E. Dale, *Morality, Rules and Consequences: A Critical Reader* (pp. 6-26). Edinburgh: Edinburgh University Press.

Shoemaker, D. (2015). *Responsibility From The Margins.* Oxford: Oxford University Press.

Sidgwick, H. (1874). *The Methods of Ethics.* London: Macmillan.

Singer, P. (1973). Famine, Affluence, and Morality. *Philosophy and Public Affairs, Vol. 1, No. 3* , 229-243.

Singer, P. (1993). *Practical Ethics, Second Edition.* Cambridge: Cambridge University Press.

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics, vol. 9, no. 3/4*, 331–352.

Singer, P., & de Lazari-Radek, K. (2017). Parfit on Objectivity and "The Profoundest Problem of Ethics". In P. Singer, *Does Anything Really Matter?: Essays on Parfit on Objectivity* (pp. 279-295). Oxford: Oxford University Press.

Slote, M. (1984). Satisficing Consequentialism. *Proceedings of the Aristotelian Society, 58*, 139–63.

Smart, J. (1961). Free-Will, Praise and Blame. *Mind, 70*, 291-306.

Smiley, M. (2017). *Collective Responsibility.* Retrieved from The Stanford Encyclopedia of Philosophy (Summer 2017 Edition): <https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility/>

Smith, M. (1994). *The Moral Problem.* Oxford: Basil Blackwell.

Sobel, D. (2016). *From Valuing to Value: A Defense of Subjectivism.* Oxford: Oxford University Press.

Spierkermannm, K. (2014). Small Impacts and Imperceptible Effects: Causing Harm with Others. *Midwest Studies in Philosophy 38 (1)*, Midwest Studies in Philosophy_ 38 (1).

Sprigge, T. (1965). A Utilitarian Reply to Dr. McCloskey. *Inquiry, 8*, 264-91.

Stawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, 48 1-25.

Stocker, M. (1976). The Schizophrenia of Modern Ethical Theories. *Journal of Philosophy 14*, 453–66.

Stocker, M. (1997). Abstract and Concrete Value: Plurality, Conflict, and Maximization. In R. Chang (ed.), *Incommensurability, Incomparability, and Practical Reason.* Cambridge: Harvard University Press.

Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy, 48*, 187–211.

Street, S. (2006). "A Darwinian Dilemma for Realist Theories of Value.". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, vol. 127, no. 1*, 109–166.

Sverdlik, S. (1987). Collective Responsibility. *Philosophical Studies, 51*, 61–76.

Taurek, J. M. (1977). Should the Numbers Count? *Philosophy and Public Affairs, 6(4)*, 293–316.

Taylor, C. (1982). The Diversity of Goods. In Sen, A., & B. Williams (eds.), *Utilitarianism and Beyond.* Cambridge: Cambridge University Press.

Thompson, D. F. (1980). Moral Responsibility of Public Officials: The Problem of Many Hands. *The American Political Science Review 74(4)*, 905-916.

Thompson, J. (1985). The Trolley Problem. *Yale Law Journal 94*, 1395-1415.

Tognazzini, N., & Coates, D. J. (2020, February 19). *Blame*. Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/fall2018/entries/blame/

Tollefsen, D. (2003). Participant Reactive Attitudes and Collective Responsibility. *Philosophical Explorations, 6*, 218-234.

Unger, P. (1996). *Living High and Letting Die.* New York: Oxford University Press.

van Ackeren, M., & Sticker, M. (2014). Kant and Moral Demandingness. *Ethical Theory and Moral Practice. 18*, 75-89.

Vargas, M. (2008). Moral Influence, Moral Responsibility. In N. Trakakis, & D. Cohen (eds.), *Essays on Free Will and Moral Responsibility* (pp. 90-122). Newcastle: Cambridge Scholars Press.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility.* Oxford: Oxford University Press.

Walla, A. P. (2015). Kant's Moral Theory and Demandingness. *Ethical Theory and Moral Practice 18 (4)*, 731-743.

Wallace, R. (1994). *Responsibility and the Moral Sentiments.* Cambridge, MA: Harvard University Press.

Williams, B. (1973). A Critique of Utilitarianism. In J. Smart, & B. Williams, *Utilitarianism: For and Against* (pp. 77-150). Cambridge: Cambridge University Press.

Williams, B. (1981). Persons, Character, and Morality. In J. Rachels (ed.), *Moral Luck: Philosophical Papers 1973–1980* (pp. 1-19). Cambridge: Cambridge University Press.

Williams, B. (1985). *Ethics and the Limits of Philosophy.* London: Fontana.

Williams, B. (1988). The Structure of Hare's Theory. In S. &. (eds), *Hare and Critics* (pp. 185-96). Oxford: Clarendon Press.

Williams, B. (1995). Acting as the Virtuous Person Acts. In R. Heinaman (ed.), *Aristotle and Moral Realism* (pp. 13-23). London: UCL Press.

Wolf, S. (2011). Blame, Italian Style. In K. a. Wallace, *Reasons and Recognition: Essay on the Philosophy of T. M. Scanlon,* (pp. 332–347). New York: Oxford University Press.

Yoder, S. D. (1998). The Nature of Ethical Expertise. *The Hastings Center Report , Vol. 28, No. 6*, 11-19.

Young, I. (2011). *Responsibility for Justice.* Oxford: Oxford University Press.