ETHICS UNDER MORAL NEUTRALITY

by

EVAN GREGG WILLIAMS

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

written under the direction of

Professor Larry Temkin

and approved by

____

____

____

____

____

New Brunswick, New Jersey

May, 2011

ABSTRACT OF THE DISSERTATION

Ethics under Moral Neutrality

By Evan Gregg Williams

Dissertation Director:
Professor Larry Temkin

How should we act when uncertain about the moral truth, or when trying to remain neutral between competing moral theories? This dissertation argues that some types of actions and policies are relatively likely to be approved by a very wide range of moral theories—even theories which have never yet been formulated, or which appear to cancel out one another's advice. For example, I argue that actions and policies which increase a moral agent's access to primary goods also tend to increase that agent's likelihood of bringing about good consequences, even under varying and mutually-incompatible hypotheses about what consequences count as "good". We therefore have a subjective, *pro tanto* moral reason to perform such actions and enact such policies—one whose justification does not require treating any particular theory as especially probable, but instead merely requires treating at least one at-least-partly consequentialist moral theory as an open hypothesis, and is therefore applicable even under conditions of moral uncertainty or moral neutrality.

My discussion begins abstractly, but as it progresses it gradually applies its framework to increasingly concrete issues. I find that the justification of some liberal

policies—in the classical sense of "liberal"—can be accomplished with significantly fewer moral assumptions than have traditionally been relied upon.

ACKNOWLEDGMENTS

First and foremost, I am profoundly grateful to my primary advisor, Larry Temkin, for his incredible patience and invaluable advice on countless drafts of this material. I also want to think the rest of my dissertation committee—Jeff McMahan, Derek Parfit, Peter Railton, and Holly Smith—for their very helpful feedback. I have benefited from more limited feedback and conversations with too many people to list, but worthy of special mention are Allen Buchanan, Gilbert Harman, Barry Loewer, Jacob Ross, Andrew Sepielli, and my father, Al Williams.

The thought process which led to this dissertation was initiated by three written works, two classic and one idiosyncratic. The classic works are John Stuart Mill's *On Liberty* (1859), arguing that some social structures are more conducive to progress than others, and John Rawls's *A Theory of Justice* (1971), specifically its concept of "primary goods"—goods which are useful regardless of one's goals. The idiosyncratic work is Eliezer Yudkowsky's *Creating Friendly AI* (2001). On the surface, his claim is that when programming an artificial intelligence which is going to be smarter and more powerful than us, we should program it to be generally "friendly" toward us and to try to figure out for itself what friendliness entails, rather than giving it specific instructions, such as "maximize human happiness", which we might regret later. However, the text can also be read metaphorically: we ethicists are the programmers, and future generations are the intelligences which will be carrying out our instructions; rather than etching our current moral values into stone, we should be instructing those future intelligences to seek out, and act upon, better or more complete values than have so far been discovered.

TABLE OF CONTENTS

INTRODUCTION

The morality of an action—by an individual or a polity—presumably supervenes on other features of that action or policy. Much theorizing has gone into the question of which features these are. Perhaps it matters whether the action displays courage. Perhaps it matters whether it treats people as equals. Perhaps it matters whether it results in more human happiness than its alternatives would have. Perhaps all of these features matter, along with a thousand others, in a complex interaction that leaves any given feature's significance altered or even reversed in some contexts. Much theorizing remains to be done.

This dissertation will not be engaging in such theorizing. Instead, it will address a different question: what should we do *now*, when the question of what makes an action moral has not yet been resolved? When we consult our best present-day moral theories to find out what to do in a given situation, only to find that they are either unable to answer, or that they offer contradictory answers, how should we proceed? Shall we be paralyzed by indecision? Shall we give up on acting morally and defer to non-moral considerations like self-interest?

I will argue that we need not be paralyzed when morally uncertain. Sometimes we do not need to know *which* features make an action morally right—in fact, sometimes we do not *need* to have the slightest clue about which features make an action morally right, although of course we normally do have some clues—to be able to see that one option is more likely to have such features than its alternatives are. If so, we have a prima facie moral reason to perform that option. It is a reason which, unlike reasons which depend for their justification on appeals to specific moral theories, does not subjectively weaken in the face of uncertainty about which moral theory is correct; it is an example of what I shall call a "theory-neutral reason". Such reasons are the topic of this dissertation.

Note that I am using the phrase "theory-neutral" in a fairly weak sense. Having a theory-neutral reason for taking an action does not require that the action *certainly* be better than its alternatives, that it be favored by *every individual* moral theory—that would be impossible, since for any action A, the theory "action A is wrong; all other actions are okay" clearly does not support that action. All it requires is that the action *probably* be better than its alternatives, that it be *on average* favored by all moral theories. This is why theory-neutral reasons are only prima facie; they can be overruled if it turns out that while an action is on average favored by all moral theories, the specific moral theories which disfavor it are—despite being less numerous—more plausible than the ones which favor it.

Although useful to everyone, or almost everyone, on occasion—for who has not, on occasion, found himself trapped in a moral dilemma and unsure where his duty lies?—this project will be of particular interest to weak moral skeptics. By "weak moral skeptics" I mean people who, unlike strong moral skeptics, accept that there might be moral truths and that these truths might be discoverable, but who, unlike non-skeptics, deny that they themselves possess information about those truths. From their perspective, all moral theories are equiprobable. Such skeptics will have previously been bereft of moral guidance, thinking "which of my options is morally right depends entirely on which moral theory is true; since I do not have any information about which moral theory is true, every option is equally likely to be morally right as every other option"—but reading my dissertation will change their situation. They will see that out of the total space of possible moral theories, some options—the ones favored by theory-neutral reasons—are supported by a wider set of theories than other options are, and no less strongly supported. If the skeptics genuinely view all moral theories as equiprobable, they will have to acknowledge that the relatively-widely-supported options are more likely to be right than the narrowly supported ones.

How can it be possible to defend the claim that one action is more likely to have right-making features than another action is, without appeal to claims about which features are right-making? I will demonstrate my method in the body of the dissertation, but for now I would like to offer an analogy to three inspirational cases from non-moral domains. The first inspiration is the scientific method. A hypothesis which has survived repeated attempts at falsification is more likely to make true predictions than a hypothesis which has never been tested—a claim which we can defend without making reference to claims about what predictions would, in fact, be true. The second inspiration is the "invisible hand" of economics. A distribution of resources which is the product of mutually-consensual exchanges in a well-functioning market is more likely to be Pareto-efficient than a distribution of resources imposed by fiat from a central authority—a claim which we can defend without making reference to claims about to whom it is most efficient to allocate each resource. The third inspiration is evolution by means of selection. An organism which is the product of selective breeding for some trait is more likely to have genes which produce that trait than an arbitrarily-chosen organism—a claim which we can defend without making reference to claims about which genes, in fact, produce the desired trait. Science, markets, and evolution all *work*; they work so well that we have based our civilization on them. We do not need to know which predictions are accurate in order to find a theory that makes those predictions; we do not need to know which allocations are accurate in order to find a distribution that includes those allocations; we do not need to know which genes code for a desired trait in order to find a genome that contains those genes; so why should we need to know which features make an action right in order to find an action that possesses those features?

I will be offering concrete examples of actions or policies which are, in cases of moral uncertainty, subjectively more likely to be morally right than their alternatives are. I do not claim that they are one hundred percent certain to be morally right. Theory-neutral reasons will generally be even more probabilistic in nature than the scientific method, the invisible hand, or evolution—all of which are themselves probabilistic to one extent or another. Sometimes the action which was subjectively most likely to be right will turn out not to have been right; theory-neutral reasons are capable of leading us astray. As a result, in cases in which we *are* confident in a particular moral theory's guidance, that theory and the moral reasons whose justifications appeal to it may well be able to undermine or outweigh our theory-neutral reasons. So this dissertation is by no means the final word in ethics; we should continue our efforts to find and apply the true moral theory. But it is important nevertheless: agents who take heed of the considerations I will be discussing will tend, on average, to do morally better than agents who ignore those considerations.

As a preview, the above claim that "we should continue our efforts to find the correct moral theory, so that we can apply that theory", while it may seem commonsensical, is itself justified by theory-neutral reasons. We do not engage in moral reflection because we are already persuaded by a given moral theory, and that theory tells us to engage in moral reflection. We engage in moral reflection because we do *not* know for certain which actions are right, and engaging in moral reflection is more likely to improve our future choices than to make them worse. Assuming that actions which have the property of "increasing the likelihood of future right actions being performed" are themselves relatively likely to be right—this is not, in fact, an assumption I want to make, but without it I could not keep the preview succinct—it follows that we have reason to reflect. "Try to find the correct moral theory" is advice that even a skeptic—again I mean a *weak* skeptic, someone who is completely uncertain about the moral truth as opposed to someone who is completely certain that ethics is a futile enterprise—could accept. I will explain the argument for moral reflection in more detail, and without the dubious assumption, when I come to it in Section 2.3.1.

Chapter One of this dissertation elaborates the definition of theory-neutral reasons and explains why they are important. Specifically, it explains why they ought to be relevant to the decision-making of anyone who has any degree of moral uncertainty, and it gives some motivation for the thought that *we* should have at least some degree of such uncertainty. It also gives a few relatively non-practical motivations for studying theory-neutral reasons—as opposed to the very practical "study theory-neutral reasons in order to apply them to your decision-making".

Chapter Two gives a basic account of how theory-neutral reasons work. It shows that some actions really are, even from a perspective of total uncertainty about the objective moral truth, subjectively better than their alternatives. Specifically, an action which meets one of the descriptions "learning about, or helping others learn about, what is morally valuable", "motivating oneself or others to pursue moral valuable ends" or "enabling oneself or others to succeed at such pursuit" should have a relatively high subjective likelihood, in comparison to actions not meeting any of those descriptions, of being instrumentally valuable. I will argue that we should expect such actions to result in intrinsically morally good consequences—and that we should expect this even if we are subjectively very unsure about what consequences qualify as intrinsically good. All else equal, this gives such actions a relatively high subjective likelihood of being objectively morally right, which gives us a subjective moral reason for taking it.

Chapter Three takes a step back and discusses decision procedures. What habits should we seek to acquire, and encourage others to acquire? What norms should we obey, and encourage others to obey? What laws and customs should we seek to institute in our society? I argue—based mainly on the finding in Chapter Two that we have a theory-neutral reason to increase people's likelihood of succeeding at their morally-motivated activities—that the "Neutral Policy", the policy most supported by theory-neutral reasons, is one which falls within the utilitarian and classical liberal camps. Many of the questions which can be asked about these theories of objective morality—for example, "when faced with the option of increasing the number of people in the world, should you try to maximize *total* utility or *average* utility or something else?"—can also be asked of the Neutral Policy. I attempt to give a detailed discussion of such questions.

Chapter Four examines some current moral and political issues from the perspective of theory-neutral reasons and the Neutral Policy: reproductive issues, such as abortion and genetic engineering, and issues involving conflicts between present people's interests and future people's interests, such as environmental conservation. Hopefully the reader will find this discussion interesting in its own right, as well as illuminating of the relatively-abstract ideas discussed in previous chapters.

Chapter Five, my concluding chapter, reviews the claims of the dissertation. It gives some quantitative examples of the kind of considerations I have in mind, for purposes of showing that theory-neutral reasons *can* be quantified—insofar as the main body of the dissertation avoids such examples, it is for the sake of not disrupting the flow of discussion, not due to any inherent vagueness in the concepts. I also examine some of the assumptions upon which my discussion rests, and discuss what would happen if those assumptions were changed.

Before beginning the discussion, I want to add a note about scope and methodology. My account is independent of theories of objective moral rightness, and of intuitions about what is objectively morally right in particular situations. It does not depend on any such theories or intuitions for its justification; it has no direct implications about which such theories or intuitions to trust; and it cannot be refuted by such theories or intuitions. So if I say something upsetting like "in a choice between saving the lives of one normal human or two severely mentally disabled humans, you have more theory-neutral reason to choose the former" and the reader firmly believes "all human lives are equally morally valuable, regardless of mental capacity", this is not necessarily a contradiction. It might well be the case that the true theory of objective morality weighs everyone's lives equally, but that people who are unaware that this theory is true have a subjective moral reason to focus on the lives of healthy people at the expense of the lives of mentally-disabled people. If the reader wants to refute my claim rather than just talking past it, he or she has two options. The first option is to reject one of my assumptions—the most important are the claim discussed in Section 1.2 that we should care about the subjective moral rightness of our actions, and the various claims defended in Section 2.2 that "moral agents", somehow defined, are *possibly* capable of learning about the moral truth, and that all else equal they usually prefer to act rightly rather than wrongly. Please note that these are assumptions about *subjective* rightness, *meta*-ethics, and *psychology* respectively; none of them are claims about what is objectively morally right. The second option is to engage my discussion on its own terms, and argue that alongside the theory-neutral reason I identify for assisting humans capable of moral agency, there is a different theory-neutral reason for assisting humans who are not. I would consider such an argument to be a friendly amendment to, not a refutation of, my view—I do not claim to be offering a *complete* list of theory-neutral reasons, only to be discussing the ones of which I am aware. What I consider to be most philosophically interesting here is the simple fact that there can *be* moral reasons whose justifications do not depend on appeals to specific moral theories, not the details of those reasons' content.

CHAPTER ONE – DEFINITION AND MOTIVATIONS

This dissertation is concerned with a particular kind of moral reason we can have for choosing some actions rather than others, one which has until now largely escaped philosophical notice. It is a reason not based on what some highly-regarded moral theory says, nor even on the overlap between several highly-regarded moral theories, but instead on an overlap between a very wide range of moral theories—so wide that it encompasses theories which appear to be simple negations of one another, and theories which nobody has even managed to formulate yet. I call the moral reasons in question "theory-neutral reasons". When other kinds of moral reasons are unavailable, theory-neutral reasons can be our default source of moral guidance; when other moral reasons *are* available, theory-neutral reasons must still be weighed into our considerations. This chapter will spell out what theory-neutral reasons are, and why the reader should be interested in them.

1.1 – Definitions

I begin with some definitions of the key terms I will be using, culminating in a definition of "theory-neutral reasons". Note that these *are* intended as definitions. I am describing the way in which *I* will be using various terms like "subjective", "reason", or "theory-neutral" in this dissertation. I am not claiming to be shedding light on how other people use such terms; it does not matter for my argument whether my usage is standard or novel. Nor am I claiming that the way I use the term is better or more natural than other ways one might use it; I will try to motivate my choice of names while explaining them, but ultimately the choice is not important to my overall argument; if the reader deems some term to be horribly inapt, he or she should feel free to mentally substitute some more apt-seeming name while reading the rest of the dissertation.

1.1.1 – The Subjective "Ought"

To begin, my domain of interest is *morality*, broadly construed so as to encompass sociopolitical questions about how groups of people should organize themselves, as well as ethical questions about how individuals should behave. My discussion will move freely between the domains of individual and collective decision-making. I take it that when making decisions, we weigh our moral beliefs against considerations such as self-interest; but I have nothing to say about the non-moral side of things. I am interested only in figuring out which direction our moral beliefs should pull.

I am also concerned with epistemic uncertainty. Whenever I use modal terms like "possible", "likely", or "uncertain" in this dissertation, I mean them in the epistemic sense of "possible/likely/uncertain *given what we know*", not in some physical, metaphysical, mathematical, or logical sense. Claims which are in some non-epistemic sense impossible can nevertheless be epistemically possible. For example, I can say that it is "likely but not certain" that Goldbach's unproven Conjecture is true and that it is "possible" that it is false, even though it is mathematically necessary that Goldbach's Conjecture have whatever truth value it has and mathematically impossible that it could have had the other. More to the point, even if—as I suspect they are—moral falsehoods are *metaphysical* impossibilities, many of them are not *epistemic* impossibilities.

Incidentally, one modal notion worthy of special note, since its modality might not be obvious, is "truth-tracking". A method of forming beliefs about a proposition P can be said to track the truth just in case it is more *likely* to arrive at a belief that P if P is true than if P is not true. Since I am understanding "likely" in an epistemic sense, I want to understand "truth-tracking" in an epistemic sense also. So, for example, if there are two ways the world could be, call them R and not-R, and if R is true then person S will believe proposition P if and only if P is true, whereas if not-R is true then S will believe P regardless of whether P is true, S will count as truth-tracking with respect to P just in case R is epistemically possible—regardless of whether R is true, and even regardless of whether R is metaphysically possible.

Putting together my concern with morality and my concern with uncertainty: I am concerned with *subjective* morality. I want to know what we can justifiably believe to be morally right, given our epistemic limitations. However, before I can discuss beliefs about what is morally right, I need to discuss moral rightness itself, so that the reader will know which beliefs I am talking about. I will begin by explicitly defining an "objective moral claim":

> An objective moral claim is a claim about what is *actually* morally right in some set of situations—not "right according to so-and-so's beliefs", "right according to such-and-such view", or "right according to such-and-such body of evidence and arguments", just right *simpliciter*.

As I am using the term, "objective" contrasts with "subjective". What makes a claim count as "objective" is what it purports to be *about*: it is a claim about how things—in this case, moral rules—really are, as opposed to being a claim about what we *believe*, or what we would have internal epistemic justification for believing given the information available to us. The word "objective" as I am using it emphatically does *not* mean "true", "well-justified", or "widely accepted", and does not contrast with "imaginary", "biased", or "idiosyncratic". Objective moral claims can be false. Even when they are true, we can lack sufficient justification for accepting them. Our confidence in them can vary from "total belief", down through "cautious acceptance", "partial credence", and "total suspension of judgment", and indeed all the way down to "total disbelief".

Since what makes a claim objective is what it is *about*, not its structure, objective moral claims can have many different structures. They can be absolute, such as "it is *always* morally wrong to kill human beings". They can be situation-specific, such as "it is morally wrong to lie *while under oath*", or even fully particular such as "*last Saturday* when your father called during dinner, it was morally wrong to hang up on him so abruptly". They can be qualified, such as "*all else equal*, it is wrong to perform actions which increase the amount of misery in the world". They can refer to levels or degrees of morality, such as "sacrificing your life for others is *supererogatory*, not obligatory" or "lying is *slightly* wrong, but killing is *severely* wrong". They can be comparative, such as "abruptly ending a conversation is *morally better than* showing up late for a meeting". They can deny the relevance of morality, such as "morality is *not relevant* to any decision whose only direct effects are on the agent himself". Objective moral claims can even make reference to facts which *would* be relevant to what is believed to be right, as long as they are not themselves claims *about* what is believed to be right; after all, facts which would be relevant to what is believed to be right are nevertheless *also* part of a complete description of the situation. For example, objective moral claims can refer to the agent's identity, as in "it is morally obligatory for *paid lifeguards* to attempt to rescue people they see drowning" or "it is morally wrong *for a Jew* to violate the commandments in the Torah". They can even refer to the agent's epistemic state, as in "it is morally wrong for an agent to assert what he *believes* to be a falsehood—i.e. to tell a lie—even if his statement is actually true", "it is morally wrong for an agent to perform an action which *might*, for all he knows, result in great harm to others, even if it will not in fact result in harm", or "it is morally wrong for an agent to violate a moral rule which he sincerely believes to be true, even if the rule in question is actually false".[1]

I shall also sometimes refer to *objective moral theories*—or sometimes just *moral theories*, since I am not sure what a "subjective moral theory" would be—which for my purposes can be defined simply as complete assignments of truth values to all objective moral claims. In general I will refer to theories by identifying their axioms, e.g. "the theory that 'actions are right insofar as they promote utility'". However, I should not be understood as tendentiously denying the view of moral particularists who think that morality cannot be finitely axiomatized. *Some* theories assign "true" to exactly one high-level principle such as "actions are right insofar as they promote utility", and what they say about all other claims can be inferred by drawing inferences from that principle. However, *other* theories are conjunctions—even infinite conjunctions—of lower-level claims, which cannot be summarized by any high-level principle. The latter still count as theories for my purposes, even if they are harder to talk about.

Exactly one objective moral theory is true. There cannot be more than one true one: since all theories are complete, we cannot have one theory that limits itself to discussing one class of situations while the other limits itself to discussing the other; which means that if they are genuinely different from one another, they must disagree about *something*. There also cannot be *no* true moral theories; if it turned out that objective morality did not exist, then the true theory would be the one I call "moral nihilism", which assigns *true* to all claims of the form "in such-and-such situations, morality is irrelevant to what the agent ought to do" and *false* to all claims which assert that some action is right, wrong, or better than another.

Some philosophers with non-realist meta-ethical views—moral non-cognitivists, some species of moral relativists, etc.—might be inclined to deny any moral theory at all, even nihilism, is true; they might argue that moral claims, even ones which make reference to an agent's attitudes when evaluating his action, cannot *have* objective truth values such as "true" and "false". Some such philosophers will have functionally equivalent notions—e.g. "reasonable" and "unreasonable"—which can play the role of truth values, in which case I hope they will keep reading and simply make the necessary substitutions. Others will lack *any* ability to evaluate objective moral claims, and so will, for my purposes, count as moral nihilists. For example, suppose than a given expressivist were willing to say "action A is right", but means by that only "I hereby express my approval of action A", which I would consider a non-moral claim. Presumably, if he were to use moral terms like "right" and "wrong" the same way *I* use those terms rather than as he wants to redefine those terms, his judgments would agree with those of nihilism—if he did not believe in nihilism, but rather thought that sometimes some actions are right and others wrong in an objective sense of "right" and "wrong", he would not have been willing to redefine "right" and "wrong" in a way that left him unable to state his belief—so I count him as a nihilist. I realize that by ignoring the distinction between various types of nihilists I am missing out on interesting meta-ethical questions, but those questions are not the subject of this dissertation.

Although I believe *that there exist* objective moral truths, I will not be taking a stand on *which* objective moral claims are true. My discussion is meant to be neutral between competing moral theories: it neither assumes, nor argues on behalf of, any specific objective moral theory or claim.

Instead, as mentioned above, this dissertation is concerned primarily with *subjective* morality. I take it that we are—at least if not led by hubris to hold beliefs with unjustifiedly-high confidence levels—uncertain about the truth value of some objective claims, including both physical claims about the world and moral claims about how we should act. Rather than wholeheartedly accepting some specific moral theory, we will assign a bit of credence to each of a great many theories. In short, we do not *know* how to act objectively rightly. This gives rise to a new question: *given* our limited evidence, the incomplete body of arguments we have considered, and our awareness of our own fallibility, what should we do? That is, for which actions could we provide the best *internal justification*—a justification appealing only to our own limited information like "it seems to me that such-and-such", not to external facts like "it really is this case that such-and-such"?

I suppose that if we had no information about which theory was correct, our credence would be smeared evenly over the whole space of possible theories—at least if we conceptualize "information" broadly enough to include judgments like "it seems like such-and-such view should be assigned a higher prior probability than its rivals" as well as more familiar types of information like "the person in front of me appears to be suffering" or "such-and-such argument seems to be sound". I choose to ignore puzzles about how to smear credences "evenly" over a set which probably has uncountably infinite cardinality, and also ignore the fact that human psychology probably would not permit the relevant credence distribution even if it were possible; this is an idealization. As it is, in light of our judgments about the intuitive plausibility of various theories and of the claims they make about particular situations, in light of the various moral arguments we have considered, and in light of whatever else can count for or against various theories, we will have higher credence in some theories than in others. In many cases our credence is only tacit; indeed, the vast majority of theories from the space of possible theories will be ones which we have never yet been formulated, never explicitly considered, but it is an open possibility, in my view deserving of at least some credence, that the true moral theory will turn out to be among this set of never-yet-considered ones.

A quick note about the word "credence": throughout this dissertation, I will be helping myself to Bayesian terminology. "Credence" in a proposition P—whether a proposition about moral claims, physical claims, or some combination thereof—represents the subjective likelihood that it is true, and ranges from 0 to 1. People can have a "conditional credence" in P-given-Q, which represents the subjective likelihood they would assign to P if they knew Q with certainty. Etc. I assume that the Bayesian picture at least *approximates* a good way to reason under conditions of uncertainty, and that therefore I will not be drifting too far astray by reasoning in that way myself.

For the most part, since this dissertation is intended to be neutral between moral theories, I will not be arguing about how our credences should be divided among the various possibilities. I do assume that it remains smeared—albeit not *evenly*—across a great many moral theories, rather than being divided among only a handful of theories or, worse, concentrated in a single theory. Section 1.4 will defend this assumption, on the grounds that none of our intuitions or arguments—especially given the fact that we are fallible beings who could have misevaluated them—are 100% conclusive, and that it would be a mistake to put all of our credence in a place which might, for all we know, be wrong. However, aside from this assumption of continued uncertainty, I will take our credences in the various moral theories as a given, and not try to argue about what those credences ought to be.

Instead, this dissertation is about the step from a hopefully-well-justified distribution of credences over moral theories, to a choice of which action to perform in a particular situation. That is, *given* that we had justification for distributing our credences over *moral theories* in some specific way, what *actions* would we have most justification for taking? The answer to this question would be obvious if we could justifiedly put absolute credence in some particular moral theory, at least if we also had access to the morally-relevant facts about our situation: in that case, we would be justified in doing whatever that theory told us to do in our situation. But when credences are divided across several moral theories, theories which will not always be in agreement with one another about what action we should perform, the correct answer becomes less obvious.

Before I discuss possible answers, I should be crystal-clear about the sense of "ought" being discussed. It is not the same "ought" that appears in objective moral claims. We might possess misleading evidence and arguments, and so assign—with perfectly good internal justification—a low credence to the objectively true theory and a high credence to one or more objectively false theories, in which case we cannot expect the action favored by the false theories to which we have assigned high credence to be the same as the action favored by the objectively true theory. Even if our evidence is not severely misleading, but only a little bit incomplete, we will still fail to assign "1" to the true theory and "0" to all false theories, and it will still be possible for the objectively right action to come apart from the subjectively right one. In fact, sometimes we know in advance that they have come apart. Consider the following case:[2]

> *The case of the burning museum:* An art museum and a neighboring fertility clinic have caught fire, endangering the original copies of some artistic masterpieces, of which reproductions exist elsewhere, as well as

> some frozen human embryos, which nobody intends to gestate. The agent's information justifies a 50% credence in a moral theory which holds that original copies of artwork are worth saving, but that frozen human embryos are unimportant when there are no plans to gestate them; and a 50% credence in a conflicting moral theory which holds that frozen human embryos are worth saving, but that original copies of artwork are unimportant when reproductions exist. The agent is leading efforts to salvage material from the burning buildings, and has three options. Option A is to concentrate salvage efforts on the museum, and will lead to the rescue of all of the artwork but none of the embryos; Option B is to concentrate salvage efforts on the fertility clinic, and will lead to the rescue of all of the embryos but none of the artwork; Option C is to salvage the easiest-to-reach material from both buildings, and will lead to the rescue of 90% of the artwork and 90% of the embryos.

Assuming that one of the two theories in which the agent is justified in placing credence is indeed correct, the objectively right option is either A or B. However, the intuitively right option for an agent in that situation to choose is C: uncertain about which type of materials should be saved, it makes sense to save as much as possible of both rather than risk focusing on the wrong one. So the agent ought—subjectively speaking—to choose C *even though he knows* that it is not the objectively right choice.

I suppose it might turn out that the true objective moral theory is "actions are objectively right or wrong to the extent that the agents performing them are subjectively justified in believing them to be right or wrong", in which case objective morality and subjective morality would not come apart. But I certainly do not want to assume that this particular theory is true: first, because that would violate my commitment to remaining neutral between objective moral theories; and second, because I do not find it terribly plausible that the *only* thing that affects an action's moral status is what moral status the agent believes it to have—that factors such as whether an action accords with a universalizable rule or whether it causes severe unnecessary pain to sentient beings are morally irrelevant whenever the agent happens to believe them to be irrelevant. So as far as I am concerned, the objective "ought" is completely distinct from the subjective "ought".

Some readers might deny that the subjective "ought" is meaningful. They might say "we ought—more or less by definition—to do whatever is objectively right, regardless of whether we know what that is". I am not entirely unsympathetic to this, since there *is* a sense in which the true objective moral theory gives the *whole* truth about morality. To whatever extent our beliefs and limited information and so on are relevant to what we ought to do, it can take them into account. Even though it may not be plausible that the objectively true theory says nothing but "actions are objectively right or wrong to the extent that the agents performing them subjectively believe them to be right or wrong", it *is* somewhat plausible that it will make allowances for cases of uncertainty about *non-moral* facts, for instance saying "maximize *expected* utility" rather than just "maximize utility". What more do we need?

What more we need is an actual decision strategy. Once we accept "we ought to do what is objectively right" and are ready to start *trying* to do what is objectively right, we have to figure out *how* to try to do what is objectively right. Perhaps this is no longer exactly a question about morality; to some extent it is just a question about instrumental rationality, a question about what strategy is most likely to fulfill our aim—an aim which happens to be something along the lines of *doing what is objectively right*. I do not really care what kind of question it is as long as the reader accepts that it is an interesting question, which I think it is: something has gone very wrong with one's motivational structure if one does *not* have the aim of doing what is objectively right, at least as one aim among others.[3] Anyhow, what we need is a strategy that maps us from *internal epistemic states* to decisions, without requiring as input any information which we do not possess. Furthermore, *we must possess the strategy itself*; the objectively true moral theory, or some "subjectivized" variant of it,[4] might be able to tell us what to do given any epistemic situation, but this does not help us if we do not know that it is the theory to which we should be paying attention.

It is not my purpose to argue for a particular strategy; instead most of the dissertation will discuss what outputs *any* adequate strategy will produce in a normal epistemic state. However, just so we have a conception of subjective rightness on the table before I turn to the more applied questions which are the focus of this dissertation, here is the strategy *I* find most plausible and will be tacitly assuming. It tells us to select the action with the highest *expected rightness*, defined like so:

> The *expected rightness* of an action with respect to a given epistemic state is the weighted average, across epistemically possible worlds, of the degree of moral rightness the action in question would have in those worlds; the description of a possible world should include an assignment of truth values to all physical and objective moral claims, and each world should be weighted in proportion to its epistemic probability.

Assuming that the epistemic probabilities of all possible worlds sum to 1 and that it makes sense to multiply "degrees of moral rightness" by scalar factors, this weighted average could be computed by finding the degree of moral rightness the action would have in each world, multiplying each of those degrees by the relevant world's epistemic probability, and then summing those products.

By "degree of moral rightness", I have in mind a scale which runs from the actions of which the moral theory stipulated to be true in a given possible world is highly disapproving, through the actions which the theory judges to be morally neutral or only just barely permissible, and on up to the actions of which the theory is highly approving. I shall sometimes refer to wrong actions as "negative" and right actions as "positive", even though there is no real requirement that the scale be numerical. Some theories may not use the entire scale: for example, a highly demanding theory might not use the positive end of the scale, instead holding that the only permissible actions are those which are morally flawless, and that there is no way for an action to be better than merely permissible.

"Expected rightness" should not look terribly unfamiliar to anyone who is accustomed to computing expected values under conditions of epistemic uncertainty. The only difference is that usually computations of expected value stipulate a theory of value and concern themselves only with factual uncertainty, whereas I am averaging across moral uncertainty as well. Even this is not a completely new idea; I am borrowing it from Ted Lockhart's *Moral Uncertainty and Its Consequences*, and those who have followed him.[5] However, Lockhart is ambiguous about one feature of the calculation which is very important for my discussion. When evaluating an action, he is not clear about whether we should instead look at the value it *does* assign to that action *in the actual world*, or whether we should look at the value it *would* assign to that action *in the epistemically possible world in which it is the true moral theory*. For example, suppose that T and U are moral theories which both endorse the claim that it is wrong to teach falsehoods and right to teach truths; suppose that we have 50% credence in theory T and 50% credence in U; and suppose that, in fact, T is the true moral theory. What is the expected rightness of teaching T? If we look at what each theory says about the action's actual value, then we would note that the action is the teaching of a truth, that both theories approve of teaching truths and so both approve of teaching T, and that therefore the action has high expected rightness. On the other hand, if we look at what each theory *would* say about the action's actual value, we see that if T were true then teaching T would be morally right, but if U were true then teaching T would be morally wrong, so the two theories oppose each other and—unless one of the theories is more vehement than the other regarding the *degree* of rightness or wrongness it assigns to teaching truths or falsehoods—teaching T has neutral expected rightness. I am explicitly using the latter approach. I care about what moral theories *would* say *if* they were true; I do not care about what they say if they are not true.

I want to temporarily set aside *practical* worries about actually implementing this procedure—our brains probably do not deal with sufficiently precise credences, and if there are sufficiently many epistemic possibilities then computing the expected rightness would take a prohibitively long time—until Section 1.2. However, there are also several unresolved *theoretical* difficulties with expanding the notion of expected rightness to encompass moral uncertainty as well as factual uncertainty. In particular, talk of "summing" degrees of moral rightness across different moral hypotheses is difficult to reconcile with the diversity of possible moral theories. Not all moral theories judge rightness in terms of a precise, scalar quantity: some only offer binary judgments of right and wrong; some only offer comparative judgments—perhaps even intransitive ones[6]—

about which options are better than which other options; some only offer vague judgments like "a little bit wrong"; some break moral rightness down into multiple components which are incommensurable with one another; and, of course, nihilism denies the distinction between "right" and "wrong" entirely. Even the theories which *do* render their judgments on a precise and linear scale may not be easily comparable with one another—for example, if one theory says "actions are right in direct proportion to the net amount of happiness they produce" and another says "actions are right in direct proportion to the net amount of justice they produce", we still need to know what the proportions in question *are* before we can sum across those theories.[7]

I think these difficulties are solvable. It seems to me that any theory which does not make reference to *how strongly* we ought to be *motivated* to perform the action it prefers in a given situation rather than the action it does not prefer—is not fully specified. It is not fully specified because there are various claims which could be added to it, such as "morality is *very* relevant to such-and-such situation; you should be absolutely unwilling to act wrongly there" or "morality is only a little bit relevant to such-and-such situation; acting rightly there isn't something on which you need to expend every ounce of your willpower", which would result in distinct theories. But if all fully-specified theories do make reference to how strongly we should be motivated to obey them in any given situation, then that motivational strength provides a common, linear scale on which to quantify their judgments, and it seems to me that most of the difficulties with expected rightness go away.

I do not want to rest too much weight on this argument. It comes worryingly close to taking a side in favor of some objective moral theories—i.e. those with well-behaved scales of evaluation—and against others, when my discussion is supposed to be neutral between all moral theories. Also, it is not at all crucial to my discussion that the reader accept an "expected rightness" approach to dealing with moral uncertainty; I will not be arguing further for such an approach. Instead I will simply speak of actions' "degree of subjective rightness". *I* am inclined to conceptualize "subjective rightness" in terms of "expected rightness", but the reader may fill it in as he or she likes. For example, if he or she prefers to say "an action's subjective rightness is simply its probability that it is not severely wrong", that will work for purposes of my discussion—although personally I dislike this approach since for any notion of "severely wrong" one can come up with a variation of the "burning museum" case in which it seems to give a bad answer. I only need three constraints on the choice of definition of "subjective rightness". The first constraint is that "subjective rightness" must depend only on epistemic states; it cannot refer to what the objective facts—including objective moral ones—*actually* are, only to how much credence we put in the various possibilities. The second constraint is that it must depend on them in the right way; it cannot treat the fact that an action is morally positive under a given open hypothesis as counting *against* the action, nor treat the fact that an action is morally negative under a given open hypothesis as counting *in favor* of the action. The third constraint is that it should be sensitive to *all* the possibilities in which we put credence. I will be assuming that if an agent obtains new information whose sole effect is to raise his credence in possible worlds in which the action would be positive while lowering his credence in possible worlds in which the action would be negative, that this will raise—albeit possibly only by a very tiny bit—the action's overall subjective rightness. This rules out options which only look at a small subset of epistemically possible worlds, such as "the subjective rightness of an action is the rightness ascribed to it by the single most plausible moral theory" or "the subjectively right action is the one favored by the single moral theory which considers the decision to be of the greatest importance". These methods are insensitive to information which raises an action's degree of rightness according to many theories but not according to the one theory on which they focus. In my opinion they are implausible methods in any case. For example, suppose that in a given situation, I am choosing between two options: Action A and Action B. Suppose I have—and am justified in having, given my available information—1% credence in each of 95 moral theories which imply "A is morally better than B, to degree D", and 5% credence in one moral theory which implies "B is morally better than A, to degree 2D". Choosing Action B would be silly. So we need a notion of subjective rightness which will be responsive to the whole range of possible moral theories and not just to the most salient one. Any notion of subjective rightness which satisfies these constraints should serve my purposes.

To conclude this section: Hume famously observes that one cannot reason *solely* from facts about what is to claims about what ought to be.[8] As far as I know, he was correct. Since this dissertation makes claims about what we ought to do—e.g. the "you ought to engage in moral reflection when morally uncertain" claim mentioned in the introduction—it has to have an "ought" premise. The premise is that we ought to do what is subjectively right. To reiterate: I do not mean this "ought"—nor any of the other "ought"s in this dissertation—in the objective moral sense; I am *not* endorsing the objective moral theory that whatever is subjectively right is also objectively right. *This dissertation does not take a stand on the question of which objective moral theory is true.* Rather I mean it in a subjective, decision-theoretic sense of "ought".

Furthermore, this premise that we ought, in some interesting sense of "ought", to do what is "subjectively right", as conceptualized within the constraints specified above, is my *only* normative premise. From it I intend to argue—appealing only to physical and meta-ethical premises—all the way down to concrete, if qualified, claims like "all else equal, and unless we possess sufficiently-strong epistemic justification for a specific moral theory that sufficiently-strongly opposes doing so, we ought—in some interesting sense of 'ought'—to sacrifice the interests of beings which are not moral agents in order to increase the primary goods available to beings which *are*". These claims come out of the structure of decision-making, not out of any commitment to a particular moral theory; they are, in that sense, morally neutral.

1.1.2 – Subjective Reasons

After Section 1.1.1, the reader has hopefully selected a notion of an action's subjective rightness with respect to an agent's epistemic state, and thinks that it matters what subjective rightness different actions possess. Unfortunately, if our epistemic credences are smeared across a very large set of physical-and-moral hypotheses, then precisely computing subjective rightness—since I have specified that the notion must be sensitive to all of those hypotheses and not just a subset of them—is going to be nearly impossible in practice. It will *not* work to actually consider each hypothesis in turn, see whether and to what extent the action we are considering would be right under that hypothesis, and then proceed to the next. For one thing, I suspect that our credence should be divided across uncountably many hypotheses, such that after we had been considering hypotheses one-by-one for an infinite length of time, there would still be infinitely more to consider. For another, I suspect—as mentioned earlier—that we should have some credence that the true moral theory is one which *nobody has yet managed to* *formulate*; evaluating an action's expected rightness with respect to such theories will be very tricky.

One might be tempted to take a shortcut. For any moral theory, we can define its *opposite*:

> To find the *opposite* of a moral theory, invert all its claims about "right" and "wrong", "better" and "worse", and so on, while retaining its claims about moral relevance and normative force.

For example, the opposite of "actions which cause non-human animals to suffer are worse, other things equal, than actions which do not" is "actions which cause non-human animals to suffer are *better*, other things equal, than actions which do not". Moral nihilism is its own opposite: since it never claims that any actions are right, wrong, or better than one another, reversing all such claims leaves it unchanged.

One might think—I will be arguing against this, but it looks plausible at first glance—that any action which is approved, to some degree, by one moral theory would be disapproved, to the same degree, by the opposite moral theory. For example, suppose that the morally salient feature of a given action is its effect on the naturalness of the environment. For any moral theory T which says that promoting a natural environment is morally positive to some degree, its opposite T* will say that promoting a natural environment is morally negative to that same degree. To the extent that the action promotes a natural environment—say, it involves setting aside wilderness areas for trees and non-human animals—then T will approve of it but T* will disapprove of it; to the extent that the action instead promotes an artificial environment—say, instead of setting aside wilderness areas, it involves bulldozing them in order to build apartment complexes for humans—then T* will approve of it but T will disapprove of it. The same will happen with any feature and any other pair of theories. So if our credences were *evenly* smeared across *all* moral theories, every action's expected rightness would—one might think—be zero, every action would have the same probability of being wrong to any degree as of being right to that degree, and so on. All actions would have the same subjective rightness and it would not matter—subjectively speaking, that is—which one we picked. All those uncountably many hypotheses, despite being open, cancel themselves out.

Our credences will not be evenly smeared across the possibility space, of course. There will be specific theories—but not any of the never-yet-formulated ones—about which we possess arguments, intuitive plausibility judgments, and so on, which raise or lower our credence in those theories relative to our credence in those theories' opposites, and perhaps also raise or lower our credence in the pair of the theories relative to the rest of the field. According to the shortcut, these specific theories are the only theories we need to look at. For example, if, because of various arguments and intuitions, we assigned 10% credence to Kantianism, 5% credence to utilitarianism, and judged that no other theories were plausible enough to stand out from the field, then to figure out the expected rightness of an action all we would have to do would be figure out how positively or negatively Kantianism judges the action—or, if we are unsure, how positively or negatively we *expect* it to judge the action—and multiply that by 10%, figure out how positively or negatively utilitarianism judges the action and multiply that by 5%, and then sum those two products. Something similar will happen with respect to notions of subjective rightness other than "expected rightness". So even though we had 85% credence in the field—85% credence that some other theory would turn out to be true—we were able to ignore that part of our epistemic state on the assumption that the field cancels itself out. We might say "Kantianism and utilitarianism are the only moral theories plausible enough to worry about". It is okay if some of the theories end up somewhat underspecified and perhaps serving as catch-alls for the portion of the field that matches an intuition about a particular case; for example, we might assign some credence to "the 'theory' that it is wrong to perform actions 'similar to' pushing the fat man in trolley cases", and then some credence to the possibility that the action under consideration is relevantly similar to pushing a fat man in a trolley case.

Under this picture of how we might *actually* make moral decisions without summing across uncountably infinite sets which include members with which we are unacquainted, we can talk about *subjective reasons* for or against an action:

> A *subjective moral reason* for taking a given action is an intellectual consideration—e.g. an argument, a bit of evidence, etc.—raising that action's subjective moral rightness above what it would otherwise have been.

For example, an argument for utilitarianism, combined with evidence that the action in question promotes utility, would constitute a subjective moral reason for taking the action.

Note that I am using "reason" here, and throughout this dissertation, in the sense of a *pro tanto* reason. There can *a* subjective moral reason in favor of a given action without the *balance* of reasons supporting it. We might have a weak argument in favor of utilitarianism and weak evidence that the action under consideration promotes utility; possessing these things would slightly raise the action's subjective rightness relative to what it would be if we did not possess them. We might also have a strong intuition in favor of Kantianism and a rigorous argument that the action violates the categorical imperative; possessing these things would greatly lower the action's subjective rightness relative to what it would be if we did not possess them. If these were the only reasons we possess, then the action will end up with negative subjective rightness; that does not mean we do not have *any* reason to perform it—we do, namely the fact that the action may promote utility and that utility may be morally significant—just that the reason in favor is outweighed by the reason against.

Note also that these are *subjective* reasons and should not be confused with the *objective* reasons which some objective moral theories include in their framework. For example, if the true moral theory includes a number of *pro tanto* moral rules which must be weighed against each other when conflicting, it would be natural to speak of the fact that an action violates one of these rules as a "reason" not to perform that action. Since my discussion is neutral between objective moral theories, I take no stand on whether the true moral theory is most naturally described as involving reasons. If it is, those are not the kind of reason which concerns me; I am concerned with subjective reasons.

I shall not be worrying about how to individuate subjective reasons. There is a sense in which an argument for utilitarianism and evidence that an action promotes utility *each* counts as a subjective reason for the action; for example, if one *already possesses* the argument for utilitarianism and then *discovers* new evidence that the action promotes utility, the subjective rightness of the action should increase as a result of that discovery. Likewise, if one already possesses evidence that the action promotes utility and then one discovers a new argument for utilitarianism, the subjective rightness of the action will also increase. So I shall feel free to regard both items—the argument for utilitarianism and the evidence that the action promotes utility—as subjective reasons for the action, even though neither would be useful without the other.

The preceding paragraph suggests two categories of subjective reason which it might be useful to distinguish. Reasons like the discovery that the action under consideration promotes utility can be called "theory-dependent reasons"; they derive their force from the elevated credence an agent has previously assigned to some specific objective moral theory. Reasons like the discovery of a new argument for utilitarianism can be called "theory-supporting reasons"; they actually *change* the agent's credences in moral theories. A morally neutral discussion like this one may not appeal to reasons of either category: if it appealed to the former, then it would be assuming objective moral claims; if it appealed to the latter, then it would be arguing on behalf of such claims. When the distinction is unnecessary, I shall refer to both categories of reason—considered individually or jointly—as a "theory-based reason".

1.1.3 – Theory-Neutral Reasons

The above picture tacitly makes an assumption which I reject: it assumes that our factual beliefs about whether a given action is supported by a given moral theory are independent of our moral beliefs about whether the given moral theory is true. This is a mistake. I will argue in Chapter Two that the two sets of beliefs should *not* be independent, and show cases in which—if the reader's credences look anything like mine—they indeed *are* not independent. For now, let us see what happens *if* they are not.

Suppose that our credence that an action A accords with theory T, conditional on T being the true theory, is higher than our credence that A accords with T, conditional on T's opposite being the true theory. That is, we have more credence in the claim "either T is true and A accords with T, or T's opposite is true and A does not accord with T" than in the claim "either T is true and A does not accord with T, or T's opposite is true and A does accord with T"; note that this is perfectly consistent with assigning the same credence to T as we assign to T's opposite. Less abstractly, suppose that in the "naturalness" case above, we expect A to promote naturalness *if* promoting naturalness is right but to promote artificiality *if* promoting artificiality is right. How this might happen will be seen in Chapter Two. For now, think about what this would mean for the calculation of subjective rightness.

If "actions are right insofar as they promote naturalness and wrong insofar as they promote artificiality" is true, then A promotes naturalness and is right. If "actions are wrong insofar as they promote naturalness and right insofar as they promote artificiality" is true, then A promotes artificiality and is right. So both theories—*despite being opposites of one another*—judge A positively. They do *not* cancel each other out; at least with respect to this pair, A should have a relatively high expected value, a relatively high chance of not being wrong, and so on.

An argument showing that an action has this feature—being relatively likely to satisfy some theory if that theory is true, but relatively unlikely to satisfy it if its opposite is true—would raise the subjective rightness of that action, and so count as a subjective reason for taking that action. What kind of reason is it? It cannot be either type of theory-based reason. It is not a theory-dependent reason, since it works even if the pair of opposing theories in question are regarded as equally plausible. It is also not a theory-supporting reason, since the new information about the action need not change our credences in the theories in question. I will call it a "theory-neutral reason":

> A *theory-neutral reason* to perform an action is a consideration that raises that action's subjective rightness despite neither constituting nor

> depending upon a consideration that raises any moral theory's subjective probability above that of its opposite.

Theory-neutral reasons are the main focus of this dissertation.

The reader will be wondering whether the name "theory-neutral" is really apt here. After all, even after we learn that A promotes naturalness if naturalness is good and artificiality if artificiality is good, there are still *some* theories which disapprove of A; for example, any theory which includes the claim "A is wrong, no matter what else is true about it". There is nothing we could learn that would make *that* theory support A. All that has been shown is that A is approved by *two* specific theories which happen to be opposites. In the more abstract case in which all I said was that "A is *relatively* likely to accord with T if T is true and *relatively* likely not to accord with T if T's opposite is true", we would not even be able to say for certain that both theories approved of A; all we could say that the *average* evaluation assigned to A by the two theories was positive. In what sense is this theory-neutral?

My idea in using the term "theory-neutral" is that a theory-neutral reason raises, or at least does not *lower*, the subjective rightness of A, no matter what our credence distribution across moral theories looks like. Returning to the naturalness case, suppose we discover, of some action A, that A is more likely than we previously thought to promote naturalness if promoting naturalness is intrinsically right and less likely than we previously thought to promote naturalness if promoting naturalness is intrinsically wrong. So the theories "actions are right insofar as they promote naturalness" and "actions are wrong insofar as they promote naturalness" both evaluate A more positively—or less negatively—than we previously thought. Meanwhile, "A is right regardless of whether it promotes naturalness" will continue to approve of A, and "A is wrong regardless of whether it promotes naturalness" will continue to disapprove of A; neither the strength of these theories' evaluations of A, nor the credence we can justifiedly place in these theories, will have been changed by the discovery that *other* theories approve of A more than previously thought. Likewise, theories like "actions are right insofar as they promote utility" will continue to evaluate A however they used to evaluate A; even if naturalness is relevant to utility, A's expected promotion of naturalness *conditional on utilitarianism being true* is unchanged by the discovery that it promotes naturalness conditional on "actions are right insofar as they promote naturalness" being true.

So under the suppositions of the case, the discovery shows that some theories approve of A more than previously thought, does not show that any theories approve of A less than previously thought, and leaves our credence distribution across theories unchanged. By the constraints I put on the otherwise free choice of definition of "subjective rightness" in Section 1.1.1—that an action's evaluation depend only on the credences, that it depend on them in the right way, and that it be responsive to all of them—it follows that no matter what credence distribution we have across moral theories, the new information cannot have lowered A's subjective rightness. Furthermore, if we place *any* credence in at least one of the opposing "naturalness" theories which support A more than we previously thought, the new information will have raised A's subjective rightness. So we can say "there are *many* credence distributions under which this discovery raises A's subjective rightness, and *none* under which it lowers A's subjective rightness". The only way a discovery could be more worthy of being called a "theory-neutral reason for A" would be if it could raise A's subjective rightness under *all* credence distributions, but this is impossible since nothing could ever change the evaluation of A under the credence distribution which assigns absolute credence to the "A is wrong regardless of what is true about it" theory. So I think it is reasonable to say that the fact that A promotes naturalness if and only if promoting naturalness is right is a "theory-neutral reason for A", rather than reserving the term "theory-neutral" for an impossibility.

Furthermore: we shall see later that, in practice, an action which is approved by one opposing pair of theories will also tend to be approved by many other opposing pairs: the same actions which promote naturalness if naturalness is good and artificiality if artificiality is good also promote challenging lives if challenging lives are good, unchallenging lives if unchallenging lives are good, and so on. Also, the usual state of affairs is for either *one* member of a pair of opposing theories to approve of any given action while the other member disapproves, or for both to regard it neutrally. So, loosely speaking, an action supported by theory-neutral reasons will be approved by *many* pairs of theories, and also by *half* of the remaining theories that do not belong to those pairs, leading the action to be supported by a *large majority* of the theories belonging to the entire field. I say "loosely speaking" because the number of possible moral theories, and of pairs of theories that support the action, may well both turn out to be infinite, and talking about proportions of infinite sets raises mathematical issues which I do not intend to address here. However, the basic point is that even though the official test for a theory-neutral reason merely requires finding *one* pair of opposing theories which both support the action in question—or even finding a pair in which one member strongly supports the action while the other only weakly opposes it—in practice we can say that an action supported by theory-neutral reasons will be an action supported by a *very wide range* of possible moral theories.

Most of this dissertation is concerned with showing that theory-neutral reasons exist and identifying their content. However, before embarking on this project I want to discuss why I think theory-neutral reasons, at least if they *do* exist and if their content *can* be identified, are interesting. That will be the task for the remainder of this chapter.

1.2 – Weighing Theory-Neutral Reasons against Others

In some special situations, theory-neutral reasons will be the only moral reasons available to an agent. One such situation would be if he assigned equal credence to every moral theory and accepted no theory-supporting reasons which might change that; in that case, theory-dependent reasons would be unavailable and so theory-neutral reasons would be the only ones of interest to him. Virtually no one is in this situation. We may admit that none of our moral arguments are perfectly rigorous and none of our moral intuitions are wholly reliable, but we will still want to give *some* weight to them, since it is possible that a non-rigorous argument could be elaborated into a sound argument, or that intuitions, even if often conflicting and unreliable, might still carry grains of truth. Another special situation would be if the agent somehow felt constrained from making use of our judgments about what moral theory is most plausible. According to some liberal traditions, this can occur in political situations.[9] Someone whose job it is to represent a diverse population should perhaps hesitate to set public policy on the grounds of theory-based reasons—reasons which will be rejected by those of his constituents who do not share his belief in the particular moral theories in question. It will be easier to justify setting public policy on theory-neutral grounds involving support from a very wide range of moral theories—remember my claim at the end of Section 1.1.3 that policies supported by theory-neutral reasons have this feature—and so can be justified without treating some theories as more worthy of being acted upon than others.

However, we usually—at least when acting as private individuals—*do* have at least some theory-based moral reasons and *are* in a position to act on them. I believe that the claim "all else equal, it is right to promote utility" is more likely to be true than the claim "all else equal, it is wrong to promote utility". I believe that the claim "all else equal, it is right to keep your promises" is more likely to be true than the claim "all else equal, it is wrong to keep your promises". And so on. Accordingly, I have theory-based subjective moral reasons for taking actions which are promotions of utility, keepings of promises, and so on. What should I do when faced with a decision in which one option is best favored by these theory-based reasons, but another option is best favored by theory-neutral reasons?

To an extent, the answer to this question is straightforward. Section 1.2 argued that we ought—in some subjective sense of "ought"—to maximize the subjective rightness of our actions. What an agent ought to do when he has many considerations relevant to subjective rightness is simply take all of them into account, figure out which of his options *does*, in fact, have the highest subjective rightness, and then choose that option.

For example, suppose that the agent is trying to decide how to vote on a project which will allow drilling for petroleum on land previously set aside as a wildlife refuge: the project, if approved, will generate some power for humans but will cause death and suffering for a number of non-human animals. Suppose that he has a theory-based reason to vote against the project: he judges that some version of "do what you can to prevent suffering and death, even of non-human animals" is somewhat likely to be the true moral theory, and that if it is the true moral theory, the suffering inflicted by the project on the animals greatly outweighs whatever small amount of human suffering could be averted by producing extra power. However, suppose that he also has a theory-neutral reason to vote for the project: it shall be shown later that many pairs of opposing theories favor actions which increase the power available to humans; whereas any theory which favors actions promoting the interests of non-human animals will have an opposite which disfavors such actions.

What this agent should do, ideally, is carefully compute subjective rightness: he should think along the lines of "well, 'prevent suffering and death, even of nonhumans' has approximately a 30% subjective likelihood of being the true moral theory, and if it were true then—notwithstanding the theory-neutral reasons in play here—voting against the project would turn out to be significantly better than voting for it; 'promote scientific progress' has a 25% likelihood, and thanks to the theory-neutral reasons in play here, if it were the true moral theory then voting for the project would turn out to be slightly better than voting against it; 'promote human social justice' has a 20% likelihood, and again thanks to the theory-neutral reasons, it too slightly favors voting for the project; and 'do not lie, cheat, or steal' has a 25% likelihood of being the true moral theory, and if it were true then both options are equally good since a vote cannot be a lie, a cheat, or a theft in the sense meant by the theory". He should use these various judgments in the way specified by the definition of subjective rightness—I still prefer the "expected rightness" calculation offered above, which will tell us to multiply 30% by "significantly better" and compare that product with the product of multiplying 25% plus 20% by "slightly better"—to arrive at a judgment about which option is subjectively best. Then he should choose that option. Sometimes theory-based reasons will outweigh theory-neutral reasons; sometimes theory-neutral reasons will outweigh theory-based reasons. It depends on how strong the various considerations are, how univocal they are—there will sometimes be theory-based reasons on both sides of the issue, cancelling each other out to an extent, and likewise there will sometimes be theory-neutral reasons on both sides, also cancelling each other out to an extent—and on what credences the agent places on the various possible moral theories.

That was the ideal. In practice, adding up considerations theory-by-theory in the manner portrayed in the previous paragraph is not going to be feasible. I can think of hundreds of possible features of actions which *might*, for all I know, be relevant to the moral status of an action that had those features. Each of these features could be given anywhere from no weight to a tremendous amount of weight, by the true moral theory. This results in a staggering number, easily in the googol range, of possible distinct moral theories—not the mere four candidates imagined above. Assigning a likelihood to so many different theories and then computing expected rightness is practically impossible. In practice what we will be tempted to do is select the handful of theories which stand out as most plausible, or most plausible in relation to their opposites, and compute expected rightness with reference only to them; even if, in aggregate, we have far more total credence in the many low-probability theories than the few high-probability ones. This strategy of ignoring the multitude of theories would work if the multitude could be relied upon to mostly cancel itself out; but when theory-neutral reasons are in play, it cannot be relied upon to do that.

Instead of summing expected rightness across a gargantuan number of theories, or across an inadequate subset of them, an alternative heuristic for identifying the option with highest subjective rightness is for an agent to sum across competing *reasons*. That is, the agent could start by assigning the two options equal estimated subjective rightness; then for each subjective reason, he could estimate how much that reason would affect subjective rightness if the ideal theory-by-theory calculation were carried out, and adjust his estimates accordingly. In the above case, an agent using this strategy would have a thought process that looks like this: "there is about a 30% subjective likelihood that the true moral theory will view the prevention of suffering and death of non-human animals as a feature which greatly increases the rightness of an action, so this raises somewhat the subjective rightness of voting 'nay'; and there is about a 75% subjective likelihood that the true moral theory will view the bringing electrical power to humans as a feature which slightly increases the rightness of an action, so this raises somewhat the estimated rightness of voting 'yea'".

The above example is still a vast oversimplification, but it does have several realistic features. The considerations I will be discussing in this dissertation tend to be relevant to an action's evaluation across a *very* wide range of possible moral theories—if anything, "75%" is low. For instance, we shall see that increasing power availability to humans promotes happiness if happiness is morally good, promotes science if science is morally good, promotes justice if justice is morally good, promotes beauty if beauty is morally good, and so on. Only the most extremely deontological theories—the ones which see morality solely as a matter of following a list of rules, none of which focus on the action's consequences rather than its nature—will be indifferent. However, the considerations I will be discussing also tend to be relatively weak and attenuated, compared to the kinds of theory-based considerations with which we are accustomed to concerning ourselves. Availability of electric power may help with the pursuit of happiness, science, justice, and so on, if they turn out to be morally good, but is not central to those pursuits the way habitats are central to animal survival. Typical theory-neutral reasons would be easily outweighed by theory-based reasons if it were not for their wider relevance across moral theories. But as it is: a large chance of a small good *can* sometimes outweigh a smaller chance of a larger good. Incidentally, I suspect the dependence on breadth of application is why theory-neutral reasons have not previously been given much attention by philosophers: utilitarians have focused on actions which *greatly* promote utility, social justice advocates have paid attention only to actions which *greatly* promote social justice, and so on. All of them were looking for the factors which mattered most according to their favorite individual moral theories; no one was looking for factors which mattered a little bit according to a very wide range of moral theories. But the latter can be as significant as the former in determining subjective rightness, if we are sufficiently humble about the present state of moral knowledge.

Incidentally, it would be remiss of me to make it sound as though theory-neutral reasons always conflict with theory-based reasons, as though we have to choose between obeying our one favorite theory or hedging our bets by doing what is a little bit right according to many possible theories. Often theory-neutral reasons will be on the same side as theory-based reasons. The features I will be discussing tend to be positive features of an action across a wide range of moral hypotheses—quite likely they will be positive features according to the reader's favorite moral theories as well, if not necessarily *very* positive. Indeed, depending on what the reader's favorite moral theory is, there may be very great overlap: I will argue in Chapter Three that theory-neutral reasons tend to favor policies which promote humanistic values while respecting individual autonomy—quite similar to what is advocated by the moral theories of liberal philosophers such as John Stuart Mill or even John Locke. By the way, do not conflate "there are theory-neutral reasons in favor of the same policies which are favored by these theories" with "there are theory-neutral reasons for believing these theories"—the first is true while the latter is utterly false. At most my discussion identifies reasons to *act as though* we believed the theories; it does not identify reasons to believe them.[10]

Anyhow, if theory-neutral reasons tend to advocate the same things as theory-based reasons, should we still care about them? I think we should. After all, assuming that a given agent is not a saint, figuring out which of his options is morally best will only be part of a given agent's decision procedure. He will also be considering other factors, such as his own self-interest. If morality and self-interest conflict, he may well start asking questions such as "is the morally best option *enough* morally better than the alternative option to justify the sacrifice I would be making by choosing it?"

If this is roughly how the agent's mind is working, a theory-neutral reason aligned with theory-based moral reasons can be every bit as important as a theory-neutral reason opposing them. "Make this sacrifice because there is a decent chance that doing so will be very morally good" is less convincing than "make this sacrifice because there is a decent chance that doing so will be very morally good, and a very high chance that it will be at least somewhat morally good". The latter might move him even if the former does not.

To review: unlike other subjective moral reasons, theory-neutral reasons are not attenuated by uncertainty about which moral theory is correct. They would remain strong all the way out to the far extreme of assigning equal credence to *all* moral theories. As a result, even a somewhat weak theory-neutral reason—i.e. one which would at most only be a tie-breaker if we knew which moral theory was true—has the ability, if we are sufficiently uncertain about which moral theory is true, to play a significant role in our decision-making, both in determining our subjective moral judgments and in determining whether it is worthwhile to obey those judgments. Theory-neutral reasons therefore deserve some attention.

1.3 – How Uncertain Are We?

I argued above that theory-neutral reasons should be taken into account alongside theory-based reasons, and can sometimes be the deciding factors in determining which available action has highest expected moral rightness. How often they will be deciding factors will depend, among other things, on how confident we are in our moral judgments. If we are very confident that we have the right moral theory, we will give a large amount of weight to theory-dependent reasons based on it, and it will be difficult for theory-neutral reasons to outweigh those claims. This section is aimed at trying to undermine such confidence. My discussion has two parts. In the first part, I will argue that our moral beliefs might be *mistaken*. Our justifications for them are not nearly as good as our justifications for mathematical or scientific beliefs. Accordingly, it is appropriate to discount theory-dependent reasons: the theories on which they depend might be false. In the second part of my discussion, I will argue that even if our moral beliefs turn out to be approximately right, they could still be *incomplete* in important ways. The goals which we believe to be moral priorities might, despite being legitimate moral goals, in fact be less important than other goals which we have not yet identified as morally significant at all. This possibility also gives us reason to discount theory-dependent reasons in comparison with theory-neutral reasons, since obeying the latter can help advance unidentified moral goals as well as identified ones.

1.3.1 – The Problem of Justification

Do we have a justification for our moral beliefs, a good reason for thinking that they are true? If they are not justified, then while they *might* be true—being unjustified is not the same as being false—we would be fools to put very much confidence in them. If they *are* justified, then we should be able to identify that justification. Can we?

I suspect that many people, asked this question, would reply "my moral beliefs are part of my religion". However, this is a non-answer. Whether, e.g., "it is wrong to have an abortion" is classified as a moral belief or as a religious belief, it still requires a justification. "I believe it on faith" is not enough; given the wide variety of mutually-incompatible religious beliefs in the world, it should be uncontroversial to say that people often believe false things on faith. "My justification is that God came to so-and-so and revealed the moral truth" is better, but still leaves open a great many questions. Are we justified in believing that so-and-so really had the experience, as opposed to being a liar, or having been misled by a dream or hallucination?[11] Even if we are convinced that the revelation occurred, are we justified in believing that so-and-so has remembered its contents correctly, and that his report, if it has been handed down through others' hands, has not been tampered with? Even if we think we have the text of the original revelation, are we justified in believing that God is benevolent and told so-and-so the truth, that God did not oversimplify for the sake of being easier to interpret, and that we *have* indeed interpreted the revelation correctly? To the best of my knowledge, there have been no divine revelations about which we can honestly answer "yes" to all of these questions.

Theology failing us, we must look to philosophy for justifications of our moral beliefs. Such attempted justifications are going to come in two broad flavors, with mixed justifications possible: deductive analysis, and induction from particular cases. I will argue that neither of these approaches is entirely satisfying.

First is the broadly deductive approach. Start with first principles and the definitions of relevant concepts, and somehow use them to identify the true moral theory. The true moral theory will tell us what to do in any situation, so we should not need empirical information about what our *actual* situation happens to be in order to identify it; so it is not insane to suppose that abstract moral truths can be uncovered by reason alone. Indeed, Kant seems to have thought that he was doing this.[12] Part of his argument may even be sound: moral reasons are distinguished from self-interested reasons by not being dependent on our individual desires; one agent is much like another when we abstract away from his individual desires; so if there is a moral law, it will be universal and apply to all agents equally, at least in its most general form; so an agent cannot consistently think that his maxim accords with the moral law when it would be undermined by being generally adopted. There are many holes here, but it seems possible that they could all be filled in, that eventually a version of this could prove sound. Of course, "the moral law is universal", even if established, would not be enough to tell us the *content* of that moral law—many different laws are capable of being followed universally.[13] Much more would need to be done.

It is not my purpose here to critique specific arguments. Instead, I would like to offer a more general comment. If conceptual analysis of ideas such as "personhood" or "rationality" were to give us knowledge of objective moral truths, it would give them to us the same way that conceptual analysis of numbers and logic gives us knowledge of objective mathematical truths—as a proof, essentially irrefutable, and justifying near certainty. Our current degree of confidence in deductive moral arguments is not yet that high. Neither Kant's Categorical Imperative, nor any other moral principle, is as impossible to rationally doubt as the Pythagorean Theorem. This can be shown easily enough: witness the people who appear to understand any given moral argument and yet continue to disagree with its conclusion. What we have at present are, *at best*, non-rigorous outlines of moral proofs rather than moral proofs themselves, leaving it entirely possible that any particular argument will turn out to have an unfixable hole. None of our arguments are good enough to warrant absolute confidence.

The lack of rigor in our deductive arguments is further suggested by people's tendency to shift to a completely different approach toward justification: the method of cases. Nobody goes around imagining triangles in an effort to find a counterexample to the Pythagorean Theorem—but we *do* go around constructing thought experiments to see how the intuitively right action accords with a proposed moral law.[14] We do it because we recognize, on some level, that our attempted deductive arguments are not successful, at least *qua* conclusive proofs, but rather are entirely fallible. Anyhow, the method of cases treats ethics more like empirical science than like mathematics: we are taking our particular reactions as data and are trying to generalize from those data.

The problem with this approach lies in the justification for trusting the "data". What is it? Of course, one can tell a story for why intuitions would be reliable: they might be the product of subconscious reasoning, or perhaps of some kind of "moral perception"; I will discuss such stories in more detail in Section 2.2.1. However, one can also offer error theories. Intuitions may be, in many or even all cases, the products of socially-inculcated prejudice, fallacious subconscious reasoning, or even a biological disposition to believe whatever helped our ancestors' genes to reproduce.[15] It is probably reasonable to let intuitions guide our actions when we have nothing better, but until we are sure about their origins, it is *not* reasonable to place high confidence in them.

Even if we did have more than sketchy arguments which might or might not flesh out, or case-by-case intuitions which might or might not be trustworthy, we would still have cause for humility. Specifically, the lack of a moral consensus is extremely troubling. Even among professional philosophers, there are Kantians, utilitarians, contractualists, virtue theorists, relativists and so on. If we possess truly persuasive justification, of any kind, for one of these moral theory, then why are so many informed and thoughtful people unpersuaded by it?[16] Perhaps the explanation is just that they have failed to appreciate our excellent justifications, due to being misled by unsound arguments or unreliable data. However, once we acknowledge that mistakes of this sort are common, we have to admit that we, too, may have made such a mistake, and adjust our confidence accordingly.[17]

Also humbling is the historical record. Consider the constraints on liberty, the impediments to progress, and the opposition to equality, which have been justified in the name of "moral decency"—this continues, in many cultures, to the present day. Consider the inquisitions, witch trials, and so on, whose perpetrators thought that torture and murder were justified as part of the great struggle against Satan. Consider the eugenics craze of the early twentieth century when it was thought, in many of the world's most advanced countries, that forced sterilization—or, in the case of Germany, extermination—of "inferiors" would better the human race. Consider the wars which have been fought, and are still fought, over competing religious or political ideologies. In short, consider *all* the horrors which have been motivated by what we now think were false moral beliefs.[18] Collectively, they form a sobering lesson on the ease with which human minds can overestimate the justification for their moral beliefs, and the harm that such overconfidence can cause. They also suggest an induction: all generations before ours have, in time, been recognized as having made serious moral mistakes; so ours probably will too.[19]

The upshot is that our moral beliefs might be wrong. I do not think we should ignore them entirely—*maybe* our arguments could be fleshed out into rigorous ones, and *maybe* our intuitions are trustworthy for some reason—but I think we must recognize a significant probability that they are leading us astray. Accordingly, we should discount subjective reasons which are dependent on the particular theories we have judged to be most credible. This raises the relative strength of theory-neutral reasons, increasing their likelihood of being the decisive factor in determining what is subjectively right.

1.3.2 – The Possibility of a Moral Catastrophe

Some readers will have a particular response in mind to all this. They will say that, yes, there are technical disagreements about the theoretical underpinnings of morality—but all of them have reasonably similar implications. Whatever the true theory turns out to be, surely that theory will agree that, for instance, we should not cause grave injury to others except in extraordinary circumstances. As for the lesson of history, well, this is indeed a special place and time, where at long last people have finally eliminated at least most of their prejudices. We—modern liberal secularists—do not want to oppress or kill *anyone*, so cannot possibly be making a mistake analogous to the ones in the above list. Our moral judgments are not *exactly* right, but surely they are at least *close* to right, able to guide us correctly in the vast majority of cases. If so, why should we worry about theory-neutral reasons? We should just follow our best guess—the moral theory which we are fairly sure is close to right—and be done with it.

The problem with this line of reasoning is that even if our "best guess" about the true moral theory *is* very close to the truth, but is just a bit incomplete—identifying all morally important considerations except one, say—the omission could still turn out to be morally catastrophic. That is, we could be participating in something morally comparable to organized slavery or the Holocaust, without even realizing that we were doing anything wrong. In that case, simply following our current best guess is unacceptable. Instead we need to find actions which at least hasten the end of the unrecognized catastrophe, and/or reduce our involvement in it. In other words, we need to find actions which will be judged at least a little bit positively by a wide range of possible moral theories.

My main strategy in this section will be to identify various moral hypotheses which I believe are open and not terribly implausible, and show how our current behavior would be not just a little bit wrong but *awful* if they are true. If these hypotheses are plausible, then, at the very least, we need to hedge our moral best guess with behavior that will at least *ameliorate* the disaster if one of these not-our-best-guess hypotheses turns out to be true. However, I will argue that they *cannot* be adequately hedged against individually by taking actions tailored to those specific hypotheses; the theory-dependent reasons suggested by one plausible hypothesis are often cancelled out by opposing theory-dependent reasons suggested by another. Instead, the only way to hedge against them will be by taking actions which tend to be at least somewhat approved by *many* moral theories, even ones which contradict one another—i.e. actions supported by theory-neutral reasons.

Perhaps a brief example of theory-neutral hedging is in order, since theory-neutral reasons have not yet been described and the reader may still be somewhat mystified about their content. I mentioned in the Introduction that moral reflection is supported by theory-neutral reasons on the grounds that it will tend to help us recognize the true moral theory, regardless of *which* moral theory this may be. No particular objective moral theory—or at least very few of them—actually says "spend time engaged in moral reflection"; instead they say things like "promote human happiness" or "do not commit murder". It would be silly for them to say "do such-and-such, and also spend time figuring out what to do". Nevertheless, when we do *not* know what to do, figuring it out is a good idea. If an unrecognized moral catastrophe is taking place, engaging in moral reflection might be the activity that *leads* us to recognize the catastrophe, after which we will put a stop to it and think "it sure was fortunate that we engaged in reflection; otherwise we would have continued to perpetrate these evils".

Further discussion of the content of theory-neutral reasons and *how* they can manage to be morally positive under a wide range of possible moral theories will have to wait for Chapter Two. For now I want to discuss the possible moral catastrophes that we need to be hedging against. Hopefully, if we knew a moral catastrophe on the scale of slavery or the Holocaust was taking place, we would already be trying to deal with it. Could one be taking place without our knowledge? I think it could. There are, broadly, two different ways in which a moral catastrophe might have escaped our attention. First would be if we failed to recognize the "catastrophe" part—falsely believing that nothing terrible was happening, e.g. due to not recognizing the victims of the catastrophe as moral subjects, or not realizing that they were being treated immorally. Second would be if we failed to recognize the "moral" part—falsely believing that *we* were not the ones doing the terrible thing, e.g. due to mistaking it for somebody else's responsibility or for an act of nature.

The most familiar example in the first category would be the abortion debate. We are quite confident that would be terrible for our society to permit a million healthy, fully-grown, innocent citizens to be killed each year by their fellow citizens; that would indeed be a Holocaust-scale event. We are also fairly confident that it would not be so terrible to permit a million human gametes to be killed each year; more than that die in a single male ejaculation. What we are not so sure about is the middle ground. Does the distinction between murder and trivial cell death occur at conception? Birth? Somewhere in between? Somewhere after birth but before adulthood? In small increments all along the process rather than at a single point? These are questions about which there is wide disagreement; existing moral theory has not resolved the issue, or at least not managed to resolve it convincingly enough to produce a consensus. Personally my "best guess" is that human lives are not morally significant until, at the earliest, the first time the humans in question become conscious. But I recognize that consciousness *itself* cannot be the relevant criterion since it is wrong to kill millions of adults *even* if it is done in their sleep, so I do not feel terribly confident about my guess. And if it turns out, contrary to my best guess, that early fetuses' lives *do* have a sufficiently-large fraction of the moral significance of adult lives, then we have a serious problem: over a million abortions *are* performed in America each year.

Many Americans are concerned enough to try to change the law and prohibit abortion, even if this involves voting for representatives whom they believe to be sub-par in other ways. Some have even gone so far as to try to assassinate abortionists or intimidate them into giving up their professions. It seems to me that such behavior could be right *if* we had sufficiently strong evidence that the crucial developmental milestone were conception. However, it is not a good hedge against the mere *possibility* that fetuses' deaths are morally significant. Pro-Lifers have rightly pointed out the collateral damage which would be caused by making legal abortions unavailable. Some women would seek out secret, amateur abortions, endangering their health and perhaps suffering punishment once the law caught up with them. Others would carry their pregnancies to term despite high costs which would have led them to seek abortion had it been available: costs ranging from a few months of discomfort and awkwardness, to more serious health complications, to effects on their lives like dropping out of school or quitting a promising career, to other social complications like damage to interpersonal relationships. We do what we can to reduce those costs, e.g. by legislating paid maternity leave for women so they do not have to quit their jobs to give birth, but the costs are still present—the million abortions that take place each year are *not* being performed for no reason. Furthermore, if abortion were unavailable and these costs were inevitable, the impersonal goal of gender equality would also be damaged: women would be susceptible to bearing these costs while men would not. Meanwhile, other pregnancies would be carried to term despite costs for people other than the mother: e.g. if the baby is so disabled that its life will be one of constant pain, or if it won't be taken care of properly, or if its birth will put such a strain on the family's resources as to damage the interests of other children.

In short, if we banned abortion—or scared off the abortionists—and it turned out that early fetal lives were *not* morally significant, the collateral damage caused by making legal abortion unavailable would *itself* constitute a moral catastrophe. In fact, I rather suspect that such a catastrophe is already taking place even without a legal ban: a million unwanted pregnancies are being aborted each year in America, but a million more are being carried to term. In many cases, the women carrying them to term are doing so despite significant reasons not to, out of a moral belief that abortion is morally terrible, or a fear of guilt and/or condemnation by others. If fetal lives are morally insignificant, then this widespread mistake is a serious disaster in its own right. But—except by trying to prevent unwanted pregnancies, which I take us to already be doing to the best of our abilities—there is no theory-dependent way to hedge against *both* the possibility that fetal lives have more significance than we think they do *and* the possibility that they have less significance than we think they do. There may, however, be theory-neutral hedges: returning yet again to the idea of reflection, seeking more information about fetuses and their moral status could be useful for helping us eventually resolve both kinds of mistakes.

The "animal rights" issue has a similar structure, except that our treatment of non-human animals is in many cases far worse than simple killing, and the number of animals being treated in the suspect way is hundreds of times larger. *Billions* of animals are living on factory farms in appalling conditions. We can try to tell stories for why their suffering matters less than otherwise-identical human suffering would—humans have a special dignity perhaps, and greater capacity for autonomy—but such stories are not entirely convincing. We can also try to hedge against the worry by avoiding the purchase of animal products and so avoiding contributing to the financial incentive to engage in such practices; but if animals matter even a thousandth as much as people do, then this is an insufficient response to the mass imprisonment, torture, and killing. We would have to institute laws restricting the rights which animal owners have over their animals. But rights reductions of this magnitude—the disruption to the economy as farms retooled for other purposes would be horrendous—are not suitable as hedges against unlikely possibilities; until we are confident that animal well-being *does* matter, we cannot justifiably perform such actions.

We might be making other mistakes about the subjects of morality. Does a person's *corpse* retain any rights—say, at least a right to respect—after the person in question is dead? The right hemispheres of right-handed people can function as independent persons if the left hemisphere is anesthetized or severed; do those hemispheres count as persons even in the case of normal human beings?[20] Is mere instantiation of the right kind of processes sufficient for consciousness, and if so are inanimate objects like computers, collective entities like nations, or even abstract entities like "the evolutionary process" capable of any sort of rudimentary consciousness and deserving of protections? Do clearly-unconscious but nevertheless symbol-laden objects like flags and books deserve respect in themselves, or may we do anything we like with them as long as we keep them secret from people who might be offended? Does mere life, even absent sentience, give plants and fungi any moral significance? If I had to guess about the answer to any individual one of these questions, my guess would be that we are not making any serious mistake; but there are enough of them—I have listed only a few—that the possibility that we are making a mistake *somewhere* is not negligible.

Even if we are right about the subjects of morality, we might be wrong about how they ought to be treated. Just to give one example, consider the euthanasia issue. When beloved household pets and companion animals become sufficiently old and sick, most of us would view it as morally mandatory to put them out of their misery. However, we keep dying *people* alive for days, weeks, or even years—even when they are in constant pain, even when they are too demented to function or maintain a shred of personal dignity, and even when they beg for death. This could easily turn out to be seriously wrong, akin to widespread torture. Undeniably it lowers the *average* well-being of the population. But euthanasia of the sufficiently-badly-off—and some euthanasia *does* take place, notwithstanding the official ban on it—might itself be akin to murder. There is no morally safe option here.

I am also concerned about how we allow people to treat *themselves*. A person who has spent decades building up his life savings can donate or gamble it away on a moment's whim, leaving his future self a pauper. He can overeat, smoke, drink, and commit other vices which risk future health consequences. If he were exposing someone *else* to such harms, we would intervene; but we allow people to do such things to themselves. This could be seriously wrong, but once again we cannot do anything about it as an individual hypothesis; we might *already* be seriously overstepping the limits of acceptable paternalism by banning many psychoactive drugs, prostitution, organ trading, and suchlike. The "organ trading" issue is particularly concerning, given that people are dying while waiting for organs to be donated without compensation. *Maybe* preventing people from commodifying themselves is a good enough reason for the restriction, but maybe it is not.

There are other familiar issues about which we might be wrong regarding what people deserve. We accept the idea that people who refuse to do work even when it is available should eventually end up homeless and hungry, notwithstanding the fact that it could be described as a kind of forced labor, because without it our economy would probably be much weaker. Our criminal justice system delicately balances a number of not-entirely-compatible goals—retribution and restitution for past crimes, prevention of future crimes, due process for people who are or could be accused of crimes, humane treatment of convicted criminals, etc.—with important rights on all sides. Our policies on issues such as affirmative action and ethnic profiling remain controversial, as we try to figure out which sorts of discrimination are acceptable and which are not. Even after we balance all relevant theory-based considerations as best we can, including theory-dependent hedging against unlikely-but-open possibilities such as "retribution does not matter at all" or "discrimination is completely unacceptable even when expected to benefit everyone", we may be still be disastrously wrong; further hedging by making use of theory-neutral reasons, if possible, is strongly desirable.

Earlier I said that there were two ways in which a moral catastrophe could go unrecognized. One would be if we failed to recognize that something terrible was occurring; some scenarios of this sort have been discussed above. The other would be if we failed to recognize our responsibility for the terrible thing that was occurring; I turn now to scenarios of this sort.

Probably the most familiar worry in this category is the problem of undeserved poverty. Many people in the world are, through no fault of their own, lacking one or more essentials: food, drinking water, shelter, sanitation, basic medical care. That this is an ongoing catastrophe is, I think, indisputable. What is not indisputable is whether this means that we are doing something wrong. Possibly the mere fact that we are standing by doing nothing when we could help—which we could, with a few minimal sacrifices—is sufficient to make our behavior count as immoral.[21] But possibly it would only count as immoral if we were somehow responsible for the problem. Are we? On the one hand, it seems that most of the impoverished people are not really impoverished because of anything *we* did as individuals—we did not personally pillage the third world—so we are not responsible in that sense. On the other hand, world poverty is not exactly a natural disaster either; part of why we are rich and they are poor does involve the history of colonialism, which means that we perhaps do owe them a debt. Also, in some countries, poverty—or the violence and oppression associated with it—is being exacerbated by market pressures from the developed world. Think of blood diamonds. Or, for that matter, think of farmers planting cash crops for rich consumers rather than staple crops for their neighbors. Even if we are only *slightly* responsible for causing the problem, or even if morality only *weakly* demands that we rescue strangers from threats which we did not cause, the fact that the problem's victims are measured in the *billions* might well be enough to make our behavior count as immoral indeed.

Can we simply help the poor, just in case? Perhaps so, but there are moral risks with giving everything one owns to some charitable institution. What if our debt is not to the poor in general but to some particular subset of the poor: the poor of our own country, or the poor of countries specifically colonized by our own country, or something like that? What if we owe a specific *kind* of aid, like help establishing good political systems? What if it turns out that we do not owe anything to the world's poor after all, and so would be giving away money that our own heirs should have been allowed to claim? Probably the safest thing to do here is to keep a substantial amount of savings in our children's college fund, and donate the rest to an assortment of different charities, in hopes that everybody with a claim on us will at least receive some small amount of payment. But make no mistake: more impoverished people will die if we make that choice than would die if we gave every last penny we owned to the single most efficient charity for curing third-world diseases. So while it may be the safest option, morally speaking, it still is not *safe*.

Another item in the category of "very bad events which we do not tend to view as our responsibility" involves risks to all of humanity. To give some examples: a collision of the Earth with a sufficiently-massive asteroid; the emergence of a super-disease combining, say, the insidiousness of AIDS with the transmissibility of the common cold; terrorist activity using world-shattering weapons; etc.[22] Even some non-extinction events like "a worldwide dictatorship arises and has high enough technology to be able to smother all hope of internal rebellion and make itself a permanent fixture" might suffice to permanently curtail the development of human civilization. While each particular disaster scenario is unlikely, the stakes are *very* high, dwarfing even the stakes of the "animal rights" worry: whether a cataclysm occurs in the near future could make the difference between human civilization collapsing within the next century or two, or the galaxy filling to the brim with a flourishing civilization lasting for many billions of years. Perhaps we should be taking action to try to decrease the likelihood of such disasters taking place and/or increase the ability of humanity to respond to them successfully if they do.

If we *knew* that one of these events was on the horizon, we would undoubtedly be making a large-scale effort—akin to the effort made during a major war—to deal with it: e.g. drafting large numbers of people into vaccine research projects, while the rest of the population practiced austerity measures to free up resources for the struggle. Should we be doing such things even when we do *not* know what threats we face? The stakes are high enough that perhaps we should. However, for the most part, the way to prepare for one future disaster is different from the way to prepare for another. Sometimes they are even directly opposed: for example, if we are worried that a tyrannical dictatorship will arise, clamp down on the population, and curtail human potential for the sake of remaining in power, we might try to strengthen civil rights like the rights to privacy and habeas corpus; but if we are worried about emerging diseases or about individual acts of terrorism, we may want to strengthen the government's ability to monitor the population and to quarantine individuals without trial, in hopes that such threats can be identified and contained before they become uncontrollable. Theory-dependent options for improving the overall future outlook, rather than just improving it with respect to one worry or another, are thus rather limited.

There are other recognizably bad events which we do not think of us our fault—non-extinction-level diseases, heartbreaks, old age, and other elements which we take for granted as part of the human condition—but which we might, in the long run, be able to make less common if we really tried. Should we try? Efforts made with the goal of improving the human condition will come at the expense of other worthwhile goals, so once again are not suitable as a hedge against the mere possibility that we ought to be engaging in this improvement. But the possibility is there nevertheless.

I have surveyed just a few possible moral catastrophes which might be taking place. The survey is by no means exhaustive. Modern morality might well be overlooking something which no one has even *thought* of worrying about, something which has not even made my list. Theory-based reasons are completely useless for hedging against this possibility: if the relevant theory has never even been formulated as a hypothesis, then we will not be able to ask how to hedge against it as an individual theory. However, theory-neutral reasons *can* be used to hedge against it; we will see that the ways in which an action can be supported by theory-neutral reasons give it such a wide support base that it can be expected to be approved not just by familiar pairs of opposing moral theories but also by pairs of opposing moral theories which nobody has yet formulated.

Section 1.4.1 argued that we do not have sufficient justification for thinking that our moral beliefs are precisely right. The present section argued that our best guess about how to behave—even if it incorporates, as best as possible given our limited knowledge, theory-dependent hedges against various specific possibilities—might still be catastrophically wrong. Hence, we should not look at theory-based reasons alone when making decisions; we should be sufficiently uncertain about our moral beliefs that theory-neutral reasons, which let us hedge against not just particular concerns but against uncertainty in general, will also play a significant role in deciding which actions are subjectively right.

1.4 – Searching for the Truth

It should be apparent from the preceding sections that the main motivation of this dissertation is an intent to take theory-neutral reasons into account during decision-making. However, I do not think this motivation is the only reason a person might be interested in theory-neutral reasons. I will argue here that theory-neutral reasons are worth exploring even for someone who does not expect them ever to be deciding factors in his decisions.

I mentioned above, and will explore somewhat in Section 3.4, the overlap between the policies supported by theory-neutral reasons and the policies supported by theory-based reasons based on liberal moral theories. In particular, we will see a theory-neutral reason for following something resembling a particular flavor of utilitarianism in our private actions, and something resembling a liberal view of rights in our treatment of each other. The similarities, while not perfect, are sufficiently strong to be surprising. Why would the actions widely suspected to be intrinsically right also turn out to be the best way to hedge against the possibility that we are morally mistaken? The coincidence cries out for an explanation.

For the reader, the most salient possible explanation of the coincidence is probably "error on the part of the Ph.D. candidate". The reader will claim that I have allowed my own favorite moral theories to corrupt my *allegedly* theory-neutral discussion. I confess to this much: I have allowed my moral views to direct my choice of topics to discuss. There are questions that could be asked about theory-neutral reasons other than the ones I am discussing. So focusing my attention where I do probably does increase the appearance of similarities between the two. But explaining the overlap in *topic* is not the same as explaining the overlap in *judgments* about that topic. I believe that my arguments are sound and do not involve illicit moral assumptions, although the reader will have to be the final judge of that.

A more interesting possibility is that the corruption runs the other direction—that we have somehow been discovering theory-neutral reasons when we thought that we were discovering theory-based moral reasons. I can tell a story of how such a thing could happen. We might have observed actions which were justified by theory-neutral reasons, recognized them on an intuitive level as morally desirable actions, but failed to consciously understand what made them morally desirable. We would then try to explain those judgments with moral theorizing that the act types in question were intrinsically morally right, even though the true explanation would be that those acts were favored by theory-neutral reasons and would therefore have turned out to be instrumentally morally right under many different moral theories. I think it would be a natural error to make.

For what it is worth—not much, given that the study method is introspection and the sample size is one—I believe that I made such an error during my own years as an impressionable young undergraduate. I was attracted to moral theories which, if accepted, would facilitate moral progress. "Surely it is morally good for moral progress to occur", I said to myself. And it surely *is* morally good for moral progress to occur—I still believe that. But I now believe that the good in question is instrumental, not intrinsic; its instrumental goodness shall be demonstrated in Section 2.3.1. As a result, in the time since I began thinking explicitly about theory-neutral reasons, my confidence in my favorite moral theories has dropped considerably. Too much of their intuitive appeal has been explained away.

The next point I want to raise is that theory-neutral considerations do not merely accord with utilitarianism or liberalism as generic abstracts. Rather, as I shall show in Chapter Three, they accord with *one specific version* of utilitarianism and *one specific version* of liberalism. This could have interesting implications for the arguments within those traditions about which versions to prefer. On one hand, if one accepts the "theory-neutral reasons explain away some of the intuitive appeal of these theories" view from the previous two paragraphs, one might think the versions which most closely accord with theory-neutral reasons are the ones which are in the most trouble. For example, one could say "I still believe in utilitarianism, but I now suspect that the intuitive appeal of *preference* utilitarianism, as opposed to hedonistic utilitarianism, was due to corruption from theory-neutral considerations". On the other hand, if one finds a less discouraging explanation for the similarities—that is, if one finds an argument for why the acts which are objectively right are also the acts which we have subjective, theory-neutral reasons to perform—then it could be *good* news for the particular versions that match. One would then say things such as "preference utilitarianism has something that other flavors of utilitarianism do not: accordance with theory-neutral considerations". Whichever way this ultimately turns out—whether accordance between objective theories and theory-neutral reasons undermine the intuitions which support those objective theories, or somehow lend additional support to them—exploring the extent of this accordance will be relevant to how confident we should be in the objective moral theories in question, and so will be a worthwhile exercise for a moral theorist.

There are also implications for more meta-theoretic questions. My methodology here is based primarily on analysis of the nature of decision-making—not on intuition pumping or psychology experiments. Assuming that my claims are correct, that the theory-neutral reasons I identify are indeed important factors in how at least some people should—at least for the subjective sense of "should" discussed in Section 1.2—behave, this will be a vindication of the analytical method. Whether a similar method would work for drawing conclusions about objective morality is an open question, but success here should justify at least a little bit of optimism.

I do not want to commit to a particular set of lessons to be drawn from my discussion here. Perhaps the reader will draw some lesson from my discussion or my conclusions which has not occurred to me; if so, great. I take myself to be identifying truths which, at least to an extent, are abstract and universal—theory-neutral reasons will apply to *any* rational beings that meet certain criteria to be spelled out in Chapter Two. There is something here to be learned, to be discovered. One does not always know in advance what implications any particular discovery would have—but that does not mean that it should be ignored. As I shall argue later in this dissertation: all else equal, knowledge is to be pursued, whether or not a present use for that knowledge is perceived. Theory-neutral reasons are part of the fabric of moral decision-making. So if we seek a more complete picture of that fabric, we should include them—whether or not we think in advance that they are important. Add this to the above points. We can study theory-neutral reasons to factor them into our decision-making, to draw lessons about moral theories, and simply because they are there to be studied. Let us now begin that study.

CHAPTER TWO – HOW THEORY-NEUTRAL REASONS ARE POSSIBLE

In Section 1.1, I offered this definition:

> A *theory-neutral reason* to perform an action is a consideration that raises that action's subjective rightness despite neither constituting nor depending upon a consideration that raises any moral theory's subjective probability above that of its opposite.

I then promised to show that there exist such reasons. This chapter aims to fulfill that promise. I will begin by reviewing the argument that theory-neutral reasons cannot exist, and showing how it might be defeated.

The argument that there can be no theory-neutral reasons goes like this. Consider any objective moral theory that says "actions with feature F are right to degree D". That theory's "opposite", as defined in Section 1.1.2, is "Actions with feature F are *wrong* to degree D". For example, "Actions are right insofar as they promote happiness" and "Actions are *wrong* insofar as they promote happiness" are opposites. Theory-neutral reasons cannot appeal to the claim that the former is more plausible than the latter; such a claim may be true but is undeniably theory-based. But without it, it seems as though these two theories would cancel one another out on any reasonable definition of subjective rightness. Either the action being evaluated does not have feature F, in which case neither theory judges it positively or negatively, or else the action being evaluated does have feature F, in which case the first theory adds some amount of subjective rightness while the second theory subtracts that same amount. Considered as a pair, it seems as though we can ignore the two theories unless we have a reason to regard one as more plausible than another. But this applies to *any* theory, so it seems as though we can conclude that there are no theory-neutral reasons. Call this the "Symmetry Argument":

> The *Symmetry Argument* attempts to show that there cannot be theory-neutral reasons because if no theory can be treated as more probable than its opposite, the pair of them will cancel each other out during the calculation of subjective rightness.

The Symmetry Argument may look sound, but it has a flaw. It assumes that what features a given action possesses are not dependent on which moral theory is true. If we can find an action A and feature F such that A is more likely to have F if a moral theory that evaluates F as a positive feature is true than if a moral theory that evaluates F as a negative feature is true, we can break the symmetry. If the theory that approves of F is true, then A will be relatively likely to have F and this likelihood will be a positive feature of A; but if the theory that disapproves of F is true, then A will be relatively unlikely to have F, and this *unlikelihood* will be a positive feature of A, so the two theories will not cancel: they will both approve of A. Another way to think about this point is that, if we individuate actions in the most fine-grained way possible, performing A while one moral theory is true is a different action from performing A while another moral theory is true. So two theories, even if they disagree about their evaluation of any *particular* action in any *particular* context, might not disagree about A.

To see how this might work, consider the following pair of objective moral theories: "it is wrong to kill a person if he has never wrongly killed anyone, and right to kill him otherwise" and "it is *right* to kill a person if he has never wrongly killed anyone, and *wrong* to kill him otherwise". These look like the sort of theories which will always cancel each other out: they both focus on the same feature, and one of them regards it positively while the other regards it negatively. The second is silly—why would killing the guilty be wrong but killing the innocent be right?—but that is beside the point. Now consider the following action: killing someone who has killed someone who had never killed anyone. If the first theory is correct, then the victim of this action has wrongly killed someone, so killing him is right. If the second theory is right, then the victim has *not* wrongly killed someone—the second theory holds that killing the innocent is not wrong—and so killing him gets evaluated as right by this theory too. Instead of canceling each other out, the two theories give overlapping judgments. They fail to cancel because whether a killing had the morally-significant property of being a killing of a "wrongful" killer depends on which moral theory was true.

Saying that we have a theory-neutral reason to kill murderers would be too fast, however. To find the opposite of a theory we are supposed to reverse *all* of its moral terms. So the correct opposite of "it is wrong to kill a person if he has never wrongly killed anyone, and right to kill him otherwise" is "it is right to kill a person if he has never *rightly* killed anyone, and wrong to kill him otherwise". These two *do* cancel one another out. If we are going to escape the Symmetry Argument, we need a theory to make a claim of the form "Actions with feature F are right to degree D", in which whether an action counts as having F *does* depend on which moral theory is true, but in which the description of F does *not* use moral language. Reminder: "depend" here should be read epistemically, as a claim about conditional probabilities. We want the epistemic probability that an action A has feature F conditional on F being objectively morally positive to be higher than the epistemic probability that A has F conditional on F being objectively morally negative.

For example, imagine that it were somehow an established fact that sprinkling holy water on a person would cause him to dissolve away to nothing, if and only if causing him to dissolve were morally right; and that otherwise it would be entirely harmless. I do not know how we could have established such a fact without already knowing something about who ought to be dissolved and having tested out the phenomenon, but imagine that we had. In such a situation, we would have a theory-neutral reason to sprinkle holy water on random people: there would be many pairs of genuinely-opposite moral theories such as "it is right to dissolve demons and wrong to dissolve witches" and "it is wrong to dissolve demons and right to dissolve witches" which would both approve of the action, on the grounds that it has a possible upside and no downside; there would be other pairs of theories, such as "it is wrong to use up holy water" and "it is right to use up holy water", which would continue to cancel themselves out; and there would be no pairs of equally-probable opposites which would disapprove of sprinkling the holy water, on net.

Unfortunately, there is no established fact in this neighborhood; holy water appears to have no special power. So while we would have a theory-neutral reason to sprinkle holy water on random people in the imaginary world in which holy water has this power to dissolve some beings, we do not have a theory-neutral reason to do so in the actual world. To find *actual* theory-neutral reasons, I need to find an action that *actually*—not just imaginably—has non-moral features which nevertheless depend on which moral theory is true. The reader is likely to be skeptical that this can be done, but this chapter will argue that it can.

2.1 – Consequences and Value

Section 2.2 will argue for the following claim: at least in general, the conditional epistemic probability of people bringing about a particular consequence given that it is a morally good consequence is higher than the conditional epistemic probability of people bringing about that same consequence given that it is not morally good. Section 2.3 will then show how we can use that fact to choose actions which make morally good consequences even *more* likely to occur. This will let us escape the Symmetry Argument. For many possible consequences C, whether the kinds of actions recommended in Section 2.3 have the feature "tending to bring about consequence C" will depend on whether the true moral theory judges this feature to be morally positive or morally negative. Before I can defend these ideas, however, I need to define some of my terms.

For my purposes, I want a broad notion of actions' "consequences":

> The *consequences* of an action are all of the facts, not indexed to the action itself, which are made true, directly or indirectly, by the action.

I realize that this is somewhat opaque, so I shall give some examples of what does and does not count as a consequence of an action.

Suppose that Adam and Bob are having an argument, and Adam shoots Bob. The consequences of this shooting are whatever we can say about the history of the universe that we could not have said if Adam had refrained from shooting Bob at that time. This includes descriptions of the shooting itself, such as "Adam kills Bob" or "a drunk man shoots a trucker at a bar on such-and-such date". It includes direct causal effects of the shooting, such as "Bob dies prematurely" or "Bob does not live to see his son's tenth birthday", as well as more remote, "but for" effects such as "Bob's son gets arrested at age eighteen for drug use, which would not have happened but for his father's absence". It includes facts that refer to events partly outside the shooting's causal radius, such as "exactly 15,312 murders take place in the United States in the year 2011" or "the money Bob borrowed five weeks before his argument with Adam is never repaid".[23]

What are not included as consequences of the shooting are facts which are indexed to the shooting. The fact that "Adam kills Bob" counts as a consequence, but the fact that "the action under discussion is Adam's killing of Bob, not some other action which eventually led to the killing" does not. The fact that "Bob dies prematurely" counts as a consequence, but the fact that "Bob's premature death is a *direct* consequence of the action under discussion, not a *remote* consequence of it" does not. Likewise, facts indexed to the agent of the shooting also do not count as consequences. The fact that "a drunk man commits a shooting on such-and-such date" counts as a consequence, but the fact that "the action under discussion was committed by someone who was too drunk to be thinking clearly about his decision" does not.

So if Adam would not have shot Bob but for Carl's perfectly-sober decision, a few years prior, to give Adam a gun for safekeeping, then it might well turn out that all of the consequences of Adam's action are also consequences of Carl's action: for example, Bob would not have died prematurely if Adam had not shot him, but Bob also—let us suppose—would not have died prematurely if Carl had not given Adam the gun, so Bob's premature death counts a consequence of both. However, many facts about Adam's action *other* than what consequences it has—for example, the fact that it was an action committed by someone who was drunk at the time—will not be true of Carl's action. For my purposes this is the essential feature of the definition of "consequences". If Action A is a consequence of Action B, then anything which is a consequence of Action A will also tend to be a consequence of Action B. I write "tend" because of course there can be complications: if Carl had not given Adam the gun for safekeeping, maybe the gun would have been stolen by Dave and used by Dave to kill Bob; if this is the way things would have gone, then while *Adam shoots Bob* is still a consequence of Carl giving Adam the gun, and *Bob dies prematurely* is still a consequence of Adam shooting Bob, *Bob dies prematurely* is not a consequence of Carl giving Adam the gun.

For my purposes, the essential feature of the above definition is that the consequences of a given action, call it A, also tend to be consequences of any action which caused A to occur—even if the type of causation involved is indirect "but for" causation. For example, if Carl gave Adam the gun used in the shooting, then—at least if we neglect the possibility that Adam might have found a gun somewhere else, which is why I wrote "tend" in the previous sentence—then Bob's premature death is a consequence *both* of Adam's decision to shoot *and* of Carl's decision to give Adam the gun. Likewise, the other consequences of the shooting are also consequences of the gun-giving, even though some of the non-consequentialist facts about the shooting, such as the fact that it *directly* resulted in Bob's premature death, may not be true of the gun-giving.

Having defined "consequences", I can now explain what I mean by "morally good".

> A given consequence would be *morally good*, according to an objective moral theory, if—and to the extent that—the theory would count the fact that an action has that consequence as a morally positive feature of the action, no matter whose action it is or what other features it has.

In other words, a consequence is morally good according to a theory if the theory, when evaluating actions which produce that consequence, evaluates them more approvingly than it would have if they did not produce that consequence, holding all else constant. Note that "morally good" here does not necessarily apply to "consequences in which many morally right actions are taking place", just to "consequences which the true moral would advocate that we produce". Of course, some objective moral theories will judge these two concepts to be co-extensive,[24] but when they diverge it is the latter one to which I mean to refer. So if we have reason to believe in a moral theory which says that Consequence C is morally good, and we also have reason to believe that Action A would produce Consequence C, then together these constitute a theory-based reason in favor of performing A. It might be outweighed by reasons against performing A, but unless it is outweighed then A is the best choice. I trust that the reader can extrapolate from this definition what I mean if I say that a consequence is "morally bad" or that one consequence is "morally better than" another. I will also sometimes use "moral goods" as a noun to refer to the consequences which are morally good.

Note that according to some moral theories, there are no such things as morally good or morally bad *consequences*. For example, consider "murder is wrong, even if it prevents many other murders; and no actions which are not murder are wrong, even if they result in many murders". It evaluates actions *only* in terms of action-relative facts—namely, whether the actions in question are murders—while denying that the actions' consequences are morally significant. I will refer to such theories as "purely non-consequentialist". Theories which are not purely non-consequentialist I will refer to as "at-least-partly consequentialist"; this includes "purely consequentialist" theories which judge the goodness of actions' consequences to be the *only* morally relevant feature of the actions, and "partly consequentialist" theories which judge the goodness of actions' consequences to be *one* morally relevant feature, but not the only one. For example, "murder is always wrong; the right action in a given situation is whichever non-murder best preserves human life" would be a partly consequentialist moral theory. It holds that consequences involving relatively more preservation of human life are better than consequences involving relatively little preservation of human life, but it also holds that some actions—namely, ones which are murders—are wrong despite having good consequences.

All of this is intended purely by way of definition. I am not trying to analyze the concepts of "consequences", "good", or "consequentialist" as they appear in everyday reasoning, in particular moral theories, or in the philosophical literature; nor have I offered any defense of such an analysis. I happen to think that my definitions *are* fairly close to established usage—which is why I chose these terms rather than others—but whether this was an apt choice is ultimately irrelevant to my argument, since my argument does not use any features of consequences or goodness other than those included within my definitions.

Incidentally, when these terms appear in intentional contexts within my discussion, they should be unpacked *de re*, not *de dicto*. For example, if I write "John thinks murders are bad", I mean this as shorthand for "John thinks that anyone, given a choice between an action whose consequences include relatively few murders occurring and an alternative action which is similar in all morally-relevant respects except that its consequences include relatively many murders occurring, should choose the former"; I do *not* mean "John would assent to the English sentence 'murders are bad'". Suppose John uses "bad" as a synonym for "wrong" rather than as the way I am using it. Suppose that if he were asked "do you think murders are bad?", he would reply "yes, I think murders are bad; no one should ever commit a murder, no matter what consequences are at stake, since murder violates the categorical imperative and actions are wrong if and only if they violate the categorical imperative". Under those circumstances, it would *not* be correct for *me* to say "John thinks murders are bad", since he does not think murders are bad in *my* sense of "bad". On the other hand, if John says "no, I do not think murders are always bad; I think they are permissible when they prevent other murders", then it might well be correct for *me* to say "John thinks murders are bad" even though John himself has just denied that murders are bad in *his* sense of "bad". What language John uses is irrelevant to what moral beliefs should be ascribed to him.

It is worth noting two facts which follow analytically from the above definitions. First: the *opposite*—see the definition in Section 1.1.2—of a purely consequentialist moral theory will also be purely consequentialist. The opposite of a partly consequentialist moral theory will also be partly consequentialist. And the opposite of a purely non-consequentialist moral theory will also be purely non-consequentialist.

Second: since the consequences of an action also tend to be consequences of whatever actions caused that action, and since the goodness of a consequence does not depend on *whose* action is being evaluated, it follows that according to any purely consequentialist moral theory, actions which cause other people to perform right actions will themselves tend to be right actions, all else equal. More generally, according to any at-least-partly consequentialist moral theory—even one which is *not* purely consequentialist—actions which cause *other* people to perform *actions which are right in virtue of bringing about good consequences* will tend to be right actions, all else equal. This is a special feature of at-least-partly consequentialist moral theories and cannot be generalized to purely non-consequentialist moral theories: a moral theory which says "generous actions are morally positive, but actions which encourage other people to act generously are morally neutral" is a perfectly coherent theory—albeit a purely non-consequentialist one.

Combining these two points: all at-least-partly consequentialist moral theories, *including pairs of opposites*, will imply that we should, all else equal, try to get people to do what they ought morally to do, at least insofar as what they ought to do is grounded in the goodness and badness of consequences. So this is an area of agreement between opposites. It may look like a rather meager area of agreement, since theories will disagree with their opposites about *what* it is that people ought morally to do. Indeed, it may look no more significant than the vacuous agreement between *all* moral theories, perforce including pairs of opposing theories, that we should do what is right and not do what is wrong, *de dicto*. But, in fact, the meager agreement about helping others fulfill *their* consequentialist duties can, with a bit of leverage, be used to break the Symmetry Argument and identify genuine theory-neutral reasons; this will be the task for the remainder of the chapter.

2.2 – Why Consequences Are More Likely to Occur if They Are Good

In his article "The Arc of the Moral Universe", Joshua Cohen defends the idea that the American South's defeat in the Civil War was foreseeable, having been made more likely by the fact that the South was fighting for a morally bad cause.[25] Part of his argument depends on a particular conception of morality and so is inadmissible here. However, another part is the claim that people "recognized" slavery as wrong and that this recognition motivated some of them to oppose it more strenuously than they otherwise would have. I find this account plausible. I also believe that similar accounts can function in hypothetical form—for example, we could claim that *if* the suffering of wild animals is bad, *then* people are more likely to put an end to it eventually than they would be if it were not bad. More abstractly: *if* any consequence is morally good, people are more likely to bring it about than they would have been if it were not morally good. Here is an outline:

> The *argument that consequences are more likely to occur if they are good* claims that all of the following features of any given consequence tends to be associated with the next:
>
> 1) How good the consequence *actually is*.
> 2) How good the consequence is *believed to be* by moral agents.
> 3) How willing moral agents are to *try to bring about* the consequence.
> 4) How likely the consequence is to *actually occur*.

I should be explicit about the definition of these four features of consequences. The first feature has already been defined in Section 2.1. A given consequence is good to the extent that actions which produce it, *in virtue of producing it*, tend to be more frequently morally right, or to be morally right to a greater average degree, than they would have been if they did not produce it. The second feature is whether people believe the consequence to have this feature. To the extent that they are inclined to view actions which they believe to produce that consequence as thereby having a relatively high degree of subjective moral rightness, they count as believing the consequence to be good. Note that this belief does not have to be occurrent; tacit belief suffices. The belief need not token the word "good"; it suffices if people make positive judgments of the actions in question. And I am averaging across the population, so a consequence's "believed goodness" goes up if more people come to regard it as good, if people who already regard it as good come to regard it as *even more* good, or if people who regard it as bad come to regard it as *less* bad. The third feature is people's average willingness, when they believe that an available action will bring about the consequence in question, to perform that action. This can also be tacit: if the opportunity to bring about the consequence in question never arises, and so the individual never thinks about whether to bring it about, that does not mean he would have been unwilling to bring it about if the opportunity had arisen. The final feature is how likely the consequence is to actually occur. This one is straightforward, so I think nothing more needs to be said about it.

The basic idea, then, is this. We can expect moral agents to be more likely to believe any given consequence to be good if it actually is good than if it is not. We can expect them to be more likely to attempt to bring it about if they believe it to be good than if they do not. And we can expect it to be more likely to come about if they are trying to bring it about than if they are not. Each of these three steps will be defended in its own section below. When they are taken together it follows, barring defeaters, that we can expect it to be more likely to come about if it actually is good than if it is not.

I want to offer some general clarificatory marks before addressing the specific connections. First, it is important to understand the claim being made. It is *not* "good consequences are more likely to be believed good, to be chosen as goals, and to come about than bad consequences are". That would be rather over-optimistic. Rather, it is "the conditional probability of a given consequence being believed good, being chosen as a goal, and coming about, given that it is good, is higher than the conditional probability of *that same* consequence being believed good, etc., given that it is not good". The distinction matters because while the goodness of consequences affects their likelihood of coming about, it is not the *only* thing which affects that likelihood. For example, if bad consequences happen to be easier to bring about than good consequences, they might well end up more likely to be brought about, in spite of their badness. I shall say more about this later.

Second, I intend the various steps of the argument, and its conclusion, to be *epistemic* claims about what conditional probabilities we should accept. They are not intended to be *causal* claims about moral ideals from Platonic Heaven somehow playing a role in the physical evolution of the universe. Nor are they intended to be *metaphysical* claims about how the world would be different if the moral truth were different—after all, while the moral truth may be epistemically uncertain, it is quite possibly metaphysically necessary. I take it that it can make sense to ask "what would the world be like if the epistemically-possible proposition P were true?" even if P is, unbeknownst to those of us for whom it is epistemically possible, metaphysically impossible. So the fact that moral truths are metaphysically necessary does *not* render "such-and-such consequence is more likely to come about if it is good than if it is not" a nonsensical claim.

Third, these claims are meant to be true *on average*, not necessarily universally. *Some* moral agents are utter fools, selfish jerks, or total incompetents. We will not necessarily want to say that *utter fools* are more likely to believe a consequence to be good if it actually is good than if it is not, that *selfish jerks* are more likely to bring about a consequence if they think it is good than if they do not, nor that *total incompetents* are more likely to bring about the consequences which they try to bring about than they would have been if aiming for something else entirely. For that matter, some consequences may also be exceptions to the rule, especially with respect to whether attempting to bring them about makes them more or less likely to actually occur. As long as the essential point still holds—that whatever consequences turn out to be morally good are at least *probably* more likely to come about than they would be if they were not morally good—the exceptions are not a problem for me.

I shall now sketch my reasons for accepting each of the three steps. I do not pretend that these sketches are absolutely conclusive, but hopefully they at least represent a good start, and place the burden of proof on those who would deny that goodness increases a consequence's likelihood of occurring.

2.2.1 – Recognition: From Actual Goodness to Believed Goodness

The first step of the argument claims that there is a connection between how much goodness a consequence *actually* has and how much it is *believed* to have. I call this step "recognition", since the claim is that people can recognize consequences' value. To put the claim another way: I assert that people's judgments about the moral goodness of outcomes are *truth-tracking*. I already mentioned this notion at the start of Section 1.1.1, but it is worth making the definition explicit:

> An agent's judgments about a particular issue are *truth-tracking* to the extent that he is more likely to believe a claim about that issue if the claim is true than he is if the claim is false.

The ideal case would be an agent who would judge a given proposition to be true if and only if it were true, and would judge it to be false if and only if it were false. Such an agent is guaranteed to be right no matter what the truth turns out to be, and we can say that his judgment about this issue is *perfectly* truth-tracking. A less paradigmatic case would be an agent who would have a 30% chance—I am still speaking in terms of the epistemic probabilities *we* assign—of judging a given proposition to be true, and a 70% chance of judging it to be false, if it were in fact true, but would have a 20% chance of judging it to be true, and an 80% chance of judging it to be false, if it were in fact false. In that case, his judgment about the issue still counts as truth-tracking—notwithstanding his evident bias toward judging it be false. Even if we strongly suspect that the proposition is true, and so forecast that this agent is more likely than not to make a mistaken judgment, he still counts as weakly truth-tracking since he is *more likely* to judge the proposition to be true if it is actually true than he is to judge it to be true if it is actually false.

In addition to being aware of the possibility that a judgment need not be perfectly truth-tracking to be slightly truth-tracking, the reader should also remember the first clarification from above: to claim that moral value judgments are at least weakly truth-tracking is *not* to claim that actually-good consequences are more likely to be regarded as good than actually-bad consequences are, only that for any *single* consequence, the conditional probability of *that consequence* being regarded as good given that it is good is greater than the conditional probability of *that same consequence* being regarded as good given that it is bad. My claim is that moral goodness is *one* feature making consequences more likely to be regarded as good; I do not claim that it is the only relevant feature, which would be ridiculous. For example, I suspect that we are predisposed to view pain as morally bad—in the above-defined sense of "something which the true moral theory would tell us to prevent, other things equal"—regardless of whether it really is morally bad, and that this one of several such predispositions which bias our moral judgments. Wishful thinking matters also—if we want a consequence to happen, we will tend to engage in rationalization, inventing false justifications for the claim that we *ought* to make it happen. Then there is historical accident: values endorsed by major world religions have a massive advantage, even if the ways in which those religions came to be popular had nothing to do with the soundness of their arguments. It in no way contradicts my position if, as a result of these effects, some moral truths are less popular than some moral falsehoods—so long as the former would be *even more* unpopular if they were false, while the latter would be *even more* popular if they were true.

What do I need to show, in order to support the premise that a consequence is more likely to be regarded as good if it is good than if it is not? First, I need it to be epistemically possible that there is such a thing as a "good consequence", in the sense being used here: a consequence which all people ought morally to bring about. Second, I need it to be epistemically possible—it need not be certain, but it needs to be possible—that at least one of the many strategies people use for forming moral beliefs is truth-tracking, in the sense of making them more likely to adopt a given moral belief if it is true than if it is false, all else equal. Third, I need it to be more likely that people use truth-tracking strategies for forming moral beliefs than that they use falsity-tracking ones.

I take the first claim to be obvious. Ethics is not such a science that we can justifiably reject claims such as "everyone should, all else equal, promote utility" or "everyone should, all else equal, promote justice" with *absolute certainty*. These claims may be true, but they are not *certainly* true.

The second claim, that at least some people are able to form truth-tracking moral beliefs, is a bit harder to defend, especially given the argument in Section 1.4.1 that familiar methods of moral reasoning might be leading us significantly astray. However, I am not contradicting myself: I suspect that our moral beliefs are, or at least might be, neither perfectly truth-tracking nor perfectly truth-independent; I suspect, rather, that they are weakly truth-tracking, that they work sometimes but not always.

Before I look at the particular methods in question, however, I want to make a general point. If one denies that people can form truth-tracking moral beliefs, then one is committed to the view that the entire enterprise of ethics is hopeless: whatever beliefs we settle on will be simply a guess, and no better a guess than if we had simply chosen moral beliefs randomly. One would have to say things such as "despite the consensus that it is wrong to torture innocent children for no purpose, that consensus was arrived at in a completely non-truth-tracking manner, and so it could just as easily be true that it is wrong to *refrain* from such torture". That is absurd. I will not, of course, be resting any weight on the particular claim that it is wrong to torture the innocent for no purpose—obvious though it is, it is still an objective moral claim and so relying on it would be non-neutral. My point is just that absolute skepticism, a denial that we can *never*, even in principle, acquire *any* information about right and wrong, is a very extreme and unattractive position, and so I am not too concerned about the possibility that my argument will be undermined by absolute skepticism.

In any case, a pragmatic argument can be made for assuming that ethics is non-hopeless. It goes like this. We want to act morally, if possible. If ethics is hopeless, then it does not matter whether we make false assumptions in our moral reasoning: whether we do our not, our beliefs will be non-truth-tracking, and whatever morally right actions we end up taking will be right by sheer chance. On the other hand, if ethics is *not* hopeless, it matters greatly whether we believe it to be hopeless. Correctly assuming that moral beliefs can be truth-tracking, and working to make ours *more* truth-tracking—and doing the other things which will be advocated in this dissertation—will make us much more likely to act rightly in the long run. So, in short, if there is any chance that moral beliefs can be truth-tracking, we have nothing to lose and much to gain by assuming that they indeed *are* truth-tracking.[26]

Of course, this pragmatic argument only works if there really is a chance that ethics is non-hopeless. If one is, pardon the oxymoron, a *dogmatic skeptic*—if one is *absolutely certain* that it is impossible to form truth-tracking moral beliefs, to the point that nothing could ever change one's mind even in principle—then one will be unmoved by the pragmatic argument. So I shall sketch some stories, which I think deserve greater than zero credence, of how at least *some* moral beliefs could turn out to be at least *slightly* truth-tracking.

Let us start with broadly analytic, deductive forms of reasoning. An agent, examining his or her favorite moral theory, and reflecting on the nature of morality, autonomy, community, and so on, could come to realize that the theory is internally inconsistent, lacks genuine normative force, arbitrarily fails to treat like cases alike, or has some other serious flaw, and that the flaw in question is not one that a correct moral theory could share.[27] I do not want to commit here to claims about what counts as a fatal flaw, but I do think that such a view is the sort of thing one could arrive at via conceptual analysis. Anyhow, the agent, moved by his reasoning, would abandon his old, flawed theory and embraces a new one which seems to lack those flaws. This will be a truth-tracking process: it will sometimes lead agents to switch from false theories to the true one, but it will *not*—at least if agents have successfully figured out what counts as a fatal flaw—lead agents to switch from the true theory to false ones. I say it will "sometimes" lead agents to switch to the true theory, because obviously this happy event will not always occur. An agent might switch from one false moral belief to another which is equally false but whose falsehood is simply less obvious. An agent might have a false moral belief but fail to identify any fatal flaws—he might even have a false moral belief in which there *are* no fatal flaws, since lacking flaws might turn out to be necessary but not sufficient for truth. This is not a problem for my account: I do not need moral reasoning to *always* work, as long as it *sometimes* works.

A seemingly bigger problem is the possibility of making an error in reasoning. A person attempting to engage in moral reasoning might be mistaken about which features of a moral theory count as fatal flaws, or about which features a given theory in fact has. Such mistaken reasoning *could* result in a switch from a true theory to a false one. But unsound reasoning does not cancel out sound reasoning: *any* theory, true or false, can be abandoned for mistaken reasons; but only false theories can be abandoned for non-mistaken reasons. So, on balance, I still think that engaging in reasoning should, on average, be expected to *increase* the accuracy of a person's moral beliefs, not to decrease it.

To illustrate the idea that abstract reasoning can be truth-tracking, consider the case of mathematics. Are our mathematical beliefs arbitrary? It seems clear that they are not. Two people independently forming beliefs about any given arithmetic problem are very likely to arrive at the same answer, and it is very likely to be the *right* answer. It seems to me that the best explanation for *how* they managed to reach the right answer instead of some arbitrary answer is that they used truth-tracking reasoning. So something like the above picture must work in at least some domains. I suspect that it works in many domains: the economic fact that "two agents with comparative advantages at producing different goods can benefit from dividing their labor and then trading with each other" and the biological fact that "given a population of organisms in a sufficiently stable environment, and a source of diversity in inheritable traits that affect reproductive fitness, the population will tend to evolve over time to be increasingly well-adapted for that environment" strike me as ultimately deriving their truth from the meaning of their words, and so as claims which *could* have been discovered analytically—even if, in our history, they were first proposed as explanations of real-world observations, rather than as abstract truths. So surely we should regard it as at least *possible* that analytic methods could yield fruit in ethics, too.

The other major strategy people tend to use during moral reasoning is appeal to intuition: take intuitions—bolstered, if applicable, by analytic arguments—about which theories are most plausible, and intuitions about what would be right for an agent to do in various specific hypothetical decision-situations, and then find a balance that conflicts with such intuitions as little as possible.[28] For example, as mentioned above, we think that it is morally wrong to torture innocent children for no purpose—and we think that it is so obviously wrong that justifying it does not require any further reasoning. This—along with other strong and clear intuitions—can serve as the data for supporting and testing moral theories: theories which imply that torturing innocent children for no purpose is permissible should be rejected.

Of course, whether this an approach is truth-tracking depends on whether our intuitions are at all reliable. As mentioned in Section 1.4.1, I have not seen any terribly convincing arguments that they are truth-tracking, and *have* seen at least some evidence that they cannot be *very* truth-tracking. However, I think they at least *might* be at least a *little* bit truth-tracking. I take it that we call a moral claim an "intuition" when we believe it without being able to spell out the reason why we believe it. Intuitions seem to spring into our minds already formed. Where do they come from? It would not surprise me if they came from a variety of different mechanisms. Maybe some of them are products of analytic reasoning, but reasoning which was conducted subconsciously. Maybe some are culturally inculcated, but nevertheless have gone through a multigenerational process of memetic evolution which somehow approximates moral reasoning or is otherwise truth-tracking: e.g. if cultures in which true moral beliefs are widespread are better able to survive than cultures in which false moral beliefs are widespread. Maybe some represent the output of some odd sort of perception.[29] Maybe some of them came about in some way I cannot even imagine, but which is truth-tracking nevertheless: "it is a mystery" is not a very satisfying explanation for why we should trust our intuition that it is wrong to torture innocent children for no purpose, but is still more satisfying than withholding belief about whether the intuition is correct.

Once again, we can look to other domains to see examples of how this strategy—appeal to intuitions—produces reliable judgments. Linguistics gives a good example. A native speaker of a given language can make intuitive judgments about which sentences are grammatical, judgments which reliably match the judgments of other native speakers, but cannot always consciously articulate the grammatical rules governing those judgments. For example, as a native English speaker I can instantly judge that "never have I heard so silly a sentence" is grammatical while "previously have I heard sillier a sentence" is not, but I find it quite challenging to explain the rules underlying the discrepancy. There *are* rules, but I do not have conscious access to them. It is at least *possible* that our intuitions of *moral* permissibility are likewise reliable, despite our inability to articulate their justifications.[30]

Of course, that it is not to say that *all* moral intuitions are reliable. While some of them may be the products of truth-tracking mechanisms, others are undoubtedly the products of non-truth-tracking mechanisms. Some could represent things we have evolved to accept instinctively, not because they are true but because accepting them increases our evolutionary fitness. Some could be the products of subconscious *mistakes* of reasoning. Some could be pure prejudice. But note that while these are examples of non-truth-tracking intuitions, they are not examples of *falsity-tracking* intuitions: a true belief is not any *more* likely to be favored by evolution or prejudice than it would if it were false, but it is also not any *less* likely to be favored; such mechanisms are completely truth-independent. So if some of our intuitions are the products of truth-tracking mechanisms and others are the products of truth-independent mechanisms, intuitions in general will still turn out to be slightly truth-tracking on average. I suppose one can tell stories about how intuitions might come to be falsity-tracking—something like "an invisible demon identifies the moral truth, and then deliberately implants contrary intuitions in us for the purposes of misleading us"—but I do not find any such stories to be plausible, or at least not any more plausible than opposing stories such as "an invisible angel identifies the moral truth, and implants good intuitions in us for the purpose of guiding us"; so they certainly cannot cancel out those intuitions which *are* truth-tracking in one of the ways suggested.

Even if we set aside both analytic reasoning *and* appeal to intuition, that does not necessarily mean that there is no connection between moral beliefs and the moral truth. I need there to be a correlation between the epistemically possible worlds in which any given moral claim is true and the ones in which it is believed. I do *not* need the direction of causation to flow from moral truths to moral beliefs; indeed, it is rather odd to speak of moral truth having causal power. It would suffice for my purposes if a more constructivist meta-ethics were true, and the relationship were more aptly described as moral truths being belief-tracking rather than moral beliefs being truth-tracking.

To show what I mean by this: I suspect that many moral claims, at least at the applied level, are established at least partly by convention. For example, society teaches its children that Promises Should Be Kept. The teaching seems to be true; saying "I promise to pay back your money" indeed places stronger moral obligations on the speaker than merely saying "I expect to pay back your money" does. But how did "society" come to know this truth, in order to teach it? It would be absurd to imagine that once upon a time, a slip of the tongue caused someone to utter the hitherto-unfamiliar phrase "I promise", that the resulting shift in moral obligations was somehow recognized by observers, and that everyone was then told about the serendipitous discovery. It had to have happened in the other order: rather than *discovering* the normative effect of the words "I promise", somebody *invented* it, and began teaching that the words "I promise" should be understood as incurring moral obligation; and then, as a *result* of people accepting this teaching, it became true that the words do incur that obligation.[31] That is, the moral claim came to be believed, and then came to be true as a result of being believed. The fact that this sort of thing can happen—assuming that we do still think that promises are morally significant, even after recognizing them as a human invention—gives us a new reason to treat moral claims as relatively likely to be believed if true and relatively likely to be disbelieved if false, other things being equal.

It would be worthwhile at this point to step back and review the overall structure of this part of my argument. I wanted to justify the assumption that at least some people's moral beliefs have better-than-chance reliability. I first argued that if it is even *possible* that they have such reliability, we have good pragmatic reasons to assume it. Then I sketched some ways in which they might be reliable: I argued that analytic reasoning about moral theories might be truth-tracking; that intuitions about particular cases might be truth-tracking; and that morality itself might be belief-tracking. This part of my argument is disjunctive; if one accepts *any* of these three options as genuine possibilities, then my argument goes through. I suspect that the whole is greater than the sum of its parts here. For example, if one denies the sort of constructivist meta-ethics under which belief-tracking moral truths are plausible, it becomes harder to deny hard-core realist meta-ethics under which moral truths are part of reality and could in principle be reflected by intuitions. And as already mentioned, denying all of them commits one to the claim that we can never form reliable judgments about moral questions—even easy-seeming questions like "should you torture innocents when it serves no purpose to do so?"

The third claim I need for the recognition premise, after the claim that there are moral truths and the claim that moral beliefs can track those truths, is the claim that moral beliefs are *more likely* on average to be truth-tracking than falsity tracking. I have already argued that analytic reasoning and appeals to intuitions should be expected to be more frequently truth-tracking than falsity-tracking; so all that remains is to argue that there is no *other*, terrible way in which people form moral beliefs and which cancels out the truth-tracking sources of belief. I do not think there is. Undoubtedly people sometimes form moral beliefs in unlicensed manners—e.g. adopting them to please peers, or as convenient excuses for behaviors that were chosen out of self-interest. But these are not falsity-tracking. A moral claim does not need to be false to be popular or convenient. Popularity, if anything, is subject to some of the same truth-tracking mechanisms discussed above; and convenience seems straightforwardly truth-independent.

That completes my argument that people sometimes recognize moral truths, or more precisely that any given moral claim is more likely to be believed by the average person if it is true than if it is false. So we should expect more people to believe a given consequence to be good, and to believe it to be more good, if it actually *is* good than if it is not. We can expect a connection between how much actual goodness a consequence has and how much goodness it is believed to have.

2.2.2 – Motivation: From Believed Goodness to Willingness of Pursuit

The next step is the one from how much goodness people believe a consequence to have and how willing they are to pursue that consequence. I call this step "motivation", since it claims that people are at least somewhat motivated by their moral judgments. The usual exceptions and clarifications apply: I do not want to claim that we never knowingly choose actions which we expect to have morally bad consequences, only that we are, on average, *less* inclined to perform such actions than we would have been if we had expected them to have morally good consequences.

I am inclined to accept an internalist picture of moral motivation, under which this step is trivial. Someone who believes that everyone, including himself, ought to act in some way, yet is not thereby inclined to act in that way, has, in my opinion, failed to understand the concept of "ought".[32] End of story. However, I know that some people will disagree with this picture, and I do not need to commit to such a view. Even if one thinks "so-and-so did such-and-such because he believed it was the right thing to do" is an incomplete explanation of an action, it should not be hard to make it a complete explanation with a small addition, such as "... and so-and-so wanted to be the kind of person who does what is right, wanted to experience pride and avoid guilt". It would absurd to think that such explanations are *never* valid.

Indeed, simple introspection rules out the possibility that moral beliefs never play a role in decision-making. *I myself* sometimes act on my moral beliefs—perform actions which I believe to be right, at least partly *because* I believe them to be right. In fact, virtually *all* of my actions have the weaker feature of being *constrained* by my moral beliefs: even if my main psychological motivation for taking a given action is that I believe it to be in my own interests, *not* that I believe it to be morally laudable, it is nevertheless true that I do not believe it to be seriously wrong, and that I *would not* be performing it if I *did* believe it to be seriously wrong.[33] For purposes of this dissertation, even actions of the latter type are sufficient to count as "morally-motivated". Presumably the reader can perform the same introspection and observe that his or her own actions are also sometimes guided, or at least constrained, by moral judgments. Therefore I think it safe to assert that people sometimes act on their moral beliefs, which is all I need for my argument; I do not care whether moral beliefs influence our behavior *on their own* or influence it *with help from other motivations*, as long as they influence it somehow or other.

Furthermore, in addition to being influenced—whether intrinsically or in combination with other desires—by their *own* moral beliefs, anybody who is remotely rational is going to pay attention to *other people's* moral beliefs as well. If one wants to make friends, attract customers, satisfy superiors, impress potential mates, or in general conduct any non-zero-sum interactions whatsoever with other human beings, it is a good idea to avoid coming across as an evildoer—which means one should try to have some idea of what they would regard as evil. So one person's belief that a consequence is good can lead, not just to *that person* trying to bring about that consequence, but to other people trying to bring it about as well.

If the reader grants my claim that people are *sometimes* motivated by moral beliefs, all that is left is for me to claim that the motivation is more often than not in the right direction—that people more often count "A is the right action" as being a consideration in *favor* of choosing A than as being a consideration *against* choosing A. I admit that occasionally people *do* count rightness as a consideration against an action. Sometimes this may be pure perversity. Other times it may be the result of people wanting to see themselves as having, or to be seen by others as having, traits associated with rule-breaking, such as independence and freedom. Of course, it is important to distinguish the people who want to be free of what they believe to be true moral rules, e.g. so that others will perceive them as tough and unpredictable and will be afraid to anger them, from the people who want to be free of what they believe to be *false* moralistic traditions, e.g. so that they can focus their energy on true values instead; I suspect the latter are more common.[34] Ultimately, this is another empirical question, but I would guess that for every gang leader who deliberately violates moral rules in order to look tough, there are thousands of jobholders, religionists, lovers, and so on, who deliberately *obey* moral rules in order to avoid offending other; and for every action performed out of sheer perversity, there are a thousand others performed out of the sincere belief that they are the right choice. Provided that my guess is correct within a few orders of magnitude, that the ratio of "people paying attention to moral beliefs—their own or others'—in order to *flout* them if the opportunity arises" to "people paying attention to moral beliefs—their own or others'—in order to *obey* them if the opportunity arises" is at least lower than one-to-one, it will be true that consequences believed to be morally good will be more likely to be pursued than they would be if they were believed to be bad.

So we have the connection between consequences' actual goodness and their believed goodness, and the connection between consequences' believed goodness and people's willingness to pursue those consequences. Does it follow that there is a connection between consequences' actual goodness and people's willing to pursue them? One could imagine defeaters: for example, suppose that the only time people were motivated to pursue consequences which were believed to be good was when those beliefs were false. Then it might be true that actual goodness was connected to believed goodness, and true that believed goodness was connected to pursuit, but still be false that actual goodness was connected to pursuit. However, no such defeater is plausible. Why in the world would someone adopt the rule "act only on false moral beliefs, not on true ones"? And how in the world would he manage to tell the difference between his false beliefs and his true ones, given that what it is to *have* a belief in the first place is to think that it is true?

So we have the claim that people are more likely to be willing to pursue a consequence if it is good than they would be if it were not good. By analogy with truth-tracking *beliefs*, we could say that such a person has value-tracking *goals*:

> A person's goals can be said to be *value-tracking* to the extent that he is more likely to pursue a consequence if it is morally good than if it is not morally good.

I trust the reader will understand what I mean if I also apply the label "value-tracking" to other propositional attitudes with the same world-to-mind direction of fit that goals have; desires, for example. Of course desires, unlike beliefs, are not necessarily *intended* to track anything: if I were to say "I want a piece of cake" and someone were to reply "you are mistaken; it is not morally good for you to have a piece of cake", I would justly object and say "I did not claim to believe that cake was morally good; I just said that I wanted it!" Notwithstanding this disanalogy, it is a useful abbreviation to be able to say that a given person's goals, desires, motivations, preferences, etc., track moral value, and I shall use this abbreviation in my later discussion. Anyhow, for now the conclusion is: on average, we should expect people's goals to be slightly value-tracking.

2.2.3 – Success: From Willingness of Pursuit to Occurrence

I turn now to the third step of the argument that consequences are more likely to occur if they are good: the claim that a particular consequence is more likely to come about if people are willing to pursue it than if they are not. I call this step "success": people at least sometimes succeed at what they try to do. For example, the likelihood of a given felon dying in the near future is higher if people want him executed than if they want his life preserved.

Saying "a consequence is more likely to come about if people are willing to pursue it than if they are not" amounts to roughly "our efforts to achieve our goals are, on average, more productive than counterproductive". I think it fairly obvious that this is true. If it were not true, our species would have gone extinct long ago: I take it that enlarged forebrains like ours are biologically very expensive, causing complications during childbirth, necessitating extra cranial protection, and consuming significant amounts of energy; what we get in exchange for these costs is an ability to choose and pursue goals; if in general pursuing goals was worse than useless at fulfilling them, this would not have been a good trade and our species would have gone extinct long ago. Of course, it might be that our efforts are frequently *ineffective*, but that is not at all the same as them being counterproductive, and not at all in tension with the claim that trying to bring about consequences makes them more, not less, likely to come about.

I do not mean to say that actions are *never* counterproductive. Sometimes people make mistakes. There may even be particular goals whose direct pursuit is on average counterproductive. For example, focusing too directly on a desire for personal happiness may be counterproductive, since happiness comes in part from being focused upon and engaged with whatever one happens to be doing.[35] Furiously striving to fall asleep, focusing closely on one's desire to forget about something, or trying to formulate a plan for acting spontaneously would also be mistakes.

However, these exceptions to the general rule of "goal-directed activity on average makes the goals in question more likely to occur" *are* exceptions. They hinge on the agent's goal involving a personal mental state different from the one he would have if he focused on his goal. Most goals do not have this structure. For example, striving to get rich is not normally counterproductive—one is more likely to get rich if one takes a high-paying job and starts saving money than if one avoids jobs and throws all one's money in a gutter. Furthermore, I think goals originating from agent-neutral moral value judgments are *particularly* unlikely to have a structure which makes them counterproductive to pursue—unlike goals based on self interest or personal duty, goals based on such judgments cannot put special emphasis on the agent's own mental state as opposed to others' mental states. If one wants to be happy, one may be better off setting this goal aside and finding something else to do; but if one wants *people in general* to be happy, I see no reason to expect direct pursuit of this goal—e.g. by trying to figure out what people need and helping them to get it—to be counterproductive. Also, if one does end up adopting some goal which cannot be pursued directly, there is some hope that one will recognize the problem and find some oblique way to pursue the goal instead. For example, *laissez-faire* economists think it is counterproductive to strive to advance the general welfare directly; but they do not think that advancing the general welfare is an impossible goal to advance, only that the correct way to advance it is via the oblique route of having everyone pursue, not the general welfare, but rather his own individual welfare.

We now have the claims that people are more likely to be willing to pursue a consequence if it is good than if it is bad, and that a consequence is more likely to be brought about if people are willing to pursue it than if they are not. Can we infer that a consequence is more likely to be brought about if it is good than if it is bad? I think we can. Defeaters would have to take the form of "efforts to bring about a consequence are much more likely to be counterproductive if we are making those efforts for moral reasons rather than non-moral ones" or "efforts to bring about a consequence are much more likely to be counterproductive if it is, in fact, a morally good consequence than if it is not". Neither of these claims strikes me as plausible. Why we chose to pursue a given goal is not especially relevant to how we will go about pursuing it, so should not be especially relevant to how successful we will be at such pursuit. So, ultimately, I think that—other things being equal—relatively good consequences can be expected to be relatively likely to be believed to be good, relatively likely to have people willing to pursue them, and, indeed, relatively likely to occur.

2.3 – Making Good Consequences More Likely to Occur

Given that there is a connection between a consequence's goodness and its likelihood of occurring, we can try to strengthen that connection. That is, we can try to make consequences *even more* likely to come about if they are good and *even less* likely to come about if they are bad. Think about the effects of actions which achieve this. If utility turns out to be good, they will make high-utility consequences more likely to occur; if equality turns out to be good, they will make high-equality consequences more likely to occur; and so on. In short, if *any* at-least-partly consequentialist moral theory is true—that is, if there *are* morally good consequences to be had—then making those consequences more likely to occur will tend to be a right action. Therefore *all* at-least-partly consequentialist moral theories, even pairs of opposites, should approve of actions which strengthen the connection between goodness and occurrence. So we have a theory-neutral reason to perform such actions—to perform actions which promote recognition, moral motivation, and success. It is a theory-neutral reason because it does *not* depend on claims about any moral theory being more probable than its opposite.

How did we escape the Symmetry Argument? Recall that the opposite of an at-least-partly consequentialist moral theory is always also an at-least-partly consequentialist moral theory. If one theory says, in whole or in part, "it is right to perform actions which cause consequence C and wrong to perform actions which avert C", its opposite will say "it is *wrong* to perform actions which cause C and *right* to perform actions which avert C". If we can manage to perform an action which has a relatively high likelihood of causing C if the former theory is true and a relatively low likelihood of causing C if the latter is true, then the two theories will *not* cancel each other out in their evaluation of it. We will have found a way to make a non-moral feature of the action—whether it causes C—contingent on moral facts, which is exactly what I claimed was needed at the start of this chapter.

We saw above that the connection between a consequence's goodness and its likelihood of occurrence is built up of three intermediate connections. The first was "recognition", linking its goodness to the goodness people believe it to have. Next was "motivation", linking people's beliefs about its goodness to their willingness to bring it about. Last was "success", linking people's willingness to bring it about to whether it actually comes about. As long as all three of these connections are already in place, strengthening any of them also strengthens the overall connection.

For example, return to the case which arose in Section 1.1, in which our credence is evenly divided between "actions are right insofar as they promote naturalness" and "actions are right insofar as they promote artificiality", among other theories. Suppose that we find some way to increase some people's likelihood of forming true moral beliefs: so those people have increased likelihood of believing "actions are right insofar as they promote naturalness" if it is true. By the argument in Section 2.2.2, this will give them an increased likelihood of trying to promote naturalness if "actions are right insofar as they promote naturalness" is true. By the argument in Section 2.2.3, this will increase the likelihood that natural consequences will actually occur if "actions are right insofar as they promote naturalness" is true. So if "actions are right insofar as they promote naturalness" is true, our action which increased people's likelihood of forming true moral beliefs will probabilistically promote naturalness and so be right. However, by the same argument *mutatis mutandis*, if "actions are right insofar as they promote artificiality" is true, our action will probabilistically promote artificiality, and so again will be right.

Likewise, in fact, for *any* moral theory of the form "actions are right insofar as they promote X", or which includes that claim among others.

So an action which strengthens any of the three connections discussed above—i.e. which promotes recognition of moral truth, motivation to act on moral beliefs, or success at achieving the goals one was pursuing—will, for any consequence C, be more likely to cause C if "actions tend to be right if they cause C" is true than if "actions tend to be wrong if they cause C" is true. This is exactly what was needed to escape the Symmetry Argument. We have a theory-neutral reason to perform such actions.

It is crucial that the reader understand that I am *not* arguing "the objective theory 'actions are right insofar as they promote recognition, motivation, and success' is more likely to be true than the objective moral theory 'actions are wrong insofar as they promote recognition, motivation, and success', and therefore we have a subjective reason to promote recognition, motivation, and success". Not only have I said nothing which supports the specific objective theory that "actions are right insofar as they promote recognition, motivation, and success", but even if I had, such an argument would be theory-based rather than theory-neutral. What I am arguing is that there are many *other* moral theories, and pairs of opposing moral theories—e.g. "actions are right insofar as they promote naturalness" and "actions are wrong insofar as they promote naturalness"—which do not even *mention* recognition, motivation, and success among their list of intrinsic moral goods, but which nevertheless deem the promotion of recognition, motivation, and success to be *instrumentally* useful. Promoting recognition, motivation, and success will be an indirect way to promote naturalness if "it is right to promote naturalness" is true; it will also be an indirect way to promote artificiality if "it is wrong to promote naturalness" is true. *That* is why I think we should do these three things; and it is a theory-neutral reason for doing them.

Note that my argument *does* require that all three connections already be in place before we try to strengthen them. More precisely, for it to be a good idea to strengthen any one of the three connections, the other two must already be in place. Otherwise strengthening one connection might *not* strengthen the overall connection. For example, if it turned out that goodness affected motivation in the wrong way—that most people, if they believed a given consequence to be morally good, were *less* likely to incorporate it into their goals than they would have been if only they had believed it to be morally bad—then improving their ability to recognize which consequences are morally good might well make good consequences *less* likely to occur rather than more. This is why I spent so much space arguing in Section 2.2 that all three elements indeed *are* already in place.

The reader might be wondering why I am phrasing everything in consequentialist terms. Instead of "increase people's ability to recognize morally good consequences, their motivation to incorporate those consequences into their goals, and their likelihood of successfully achieving those goals", why not make the broader suggestion "increase people's ability to recognize moral truths, their motivation to incorporate those truths into their decisions, and their likelihood of successfully doing what they try to do" and capture deontological and virtue-theoretic moral codes as well?

The reason I cannot do that is because my argument relies on the special feature of consequences remarked upon in Section 2.1: the fact that if I cause someone else to cause good consequences, I have thereby caused good consequences and so done the right thing according to consequentialism. On the other hand, it is simply not the case that if I cause someone else to obey his non-consequentialist duties, I have necessarily obeyed my own non-consequentialist duties; maybe I have, maybe I have not. Even if we all have identical duties, which is by no means guaranteed, helping him fulfill his might conflict with me fulfilling mine: imagine one child loudly reminding another that they are both supposed to be quiet; this child is helping the other child remember her obligation, but is not obeying his own identical obligation. So "it is right to increase people's ability to recognize and obey non-consequentialist moral rules" depends on the objective moral claim that it is morally good for people to obey their duties—i.e. "it is right to help others fulfill their duties". If every moral theory were regarded as no more probable than its opposite, "it is right to help others fulfill their duties" would be cancelled out by "it is *wrong* to help others fulfill their duties". Likewise for specific duties such as a duty not to kill; "it is right to prevent others from killing" would be cancelled by "it is wrong to prevent others from killing".

Contrast this situation with the tautological claim "it is right to help others bring about good consequences". Reversing all moral terms in it yields the equally-tautological "it is *wrong* to help others bring about *bad* consequences", which does *not* cancel the former. Similarly if we consider a specific good consequence such as utility. The opposite of "utility is good; i.e. it is right to bring it about directly and also right to help others bring it about" is "utility is bad; i.e. it is wrong to bring it about directly and also wrong to help others bring it about". Both of these *agree* that it is right to help others bring about what is good; they only disagree about whether utility is good.

So the theory-neutral reasons I will identify in this section are essentially consequentialist. They are consequentialist in structure—telling us to promote recognition, moral motivation, and success as instrumentally good consequences. More importantly, they are consequentialist in justification—they overcome the Symmetry Argument by being endorsed by *consequentialist* moral theories, or at least by the consequentialist *parts* of moral theories, and their opposites. So they are dependent on there being *some* at-least-partly consequentialist moral theories among the live moral hypotheses, but not on any *particular* moral theory being more probable than its opposite. A genuinely theory-dependent reason is undermined by moral uncertainty—the more spread out our credences are, the less credence will rest on any particular moral theory. Theory-neutral reasons are not so undermined. At least if we speak loosely and neglect problems involving proportions of infinite sets, we can say that in the space of possible moral theories, *most* incorporate at least *some* claims about good or bad consequences. For example, for every purely non-consequentialist theory of the form "there is a side constraint against actions of type X", there is a partly-consequentialist theory of the form "there is a side-constraint against actions of type X, but within that side constraint we should promote consequence C", another that says "there is a side-constraint against actions of type X, but within that side constraint we should promote consequence D", and so on. The more evenly our credences are spread across this space, the more plausible we will have to regard the claim "there are such things as morally good consequences", and so the more reason we will have for trying to make morally good consequences occur. Theory-neutral reasons are undermined not by uncertainty but by some types of confidence: it is only if one were justifiedly confident that a purely non-consequentialist moral theory is true that one would be licensed in ignoring opportunities to promote recognition, motivation, and success.

I do not think any such confidence could be justified. Given a decision in which it is known that one option would result in a world in which happiness, justice, progress, and glory abound, in which people possess generosity, honor, wisdom, and courage, and treat each other with kindness, fairness, honesty, and respect; and in which it is also known that the alternative option would result in a world of suffering, injustice, ignorance, and inhumanity, I find it difficult to imagine that such facts even *could* all be irrelevant. Being *confident* that they are all irrelevant would be even more difficult. However, this *is* a weak point in my argument. I cannot, without appealing to these intuitions about which theories are plausible, say "theory-neutral reasons for an action will raise its subjective rightness under *any* credence distribution across theories". All I can say is "theory-neutral reasons for an action—at least, those presently under discussion; see Section 2.4.1 for one which is slightly more broadly applicable—will raise its subjective rightness under any credence distribution across theories *which assigns at least some credence to at least one at-least-partly consequentialist theory*; they will neither raise nor lower the action's subjective rightness under credence distributions which assign credence only to purely non-consequentialist theories". I feel that it is fair to call these reasons "theory-neutral" despite their dependence on some at-least-partly consequentialist theory receiving at least a little bit of credence; certainly they still fit the definition of "theory-neutrality" offered in Section 1.1.

A final caveat is that the reasons I am discussing are at best *pro tanto* reasons. If one is considering an action which makes good consequences, in the abstract, more likely to occur, but which also has some other feature—e.g. *being a lie* or *causing pain*—which one has a theory-based reason to believe to be morally negative, the theory-based reason against the action might well outweigh the theory-neutral reason in favor of the action. Similarly, if one has evidence that a particular person's goals are *not* based on a truth-tracking moral judgment—for example, one happens to know that he is being motivated by selfishness, and would be pursuing the same goals no matter what moral theory was true—this undermines the theory-neutral reason to help that particular person fulfill those goals.

In this context, it is also worth hearkening back to my earlier warnings about what exactly I have argued for and what I have not. For example, I argued that the conditional probability of any given consequence being believed good if it really is good is higher than the conditional probability of *that same* consequence being believed good if it really is bad, but I did not argue that the probability of the consequences which really are good being believed good is higher than the probability of the *other* consequences which really are bad being believed good. So it is fully consistent with everything I have said for it to be the case that the most popular moral beliefs are seriously wrong. If one has reason to suspect that this is the case—i.e. one knows which moral beliefs are most common and one has reason to believe that those particular beliefs are seriously wrong—then one has a theory-based reason not to strengthen the other two connections. One would still *have* the theory-neutral reason to strengthen those connections—it would still be true that, if one were to set aside all information about which theories are more plausible than their opposites, strengthening those connections would make good consequences more likely to occur—but it would be outweighed by the theory-based reason not to strengthen them.

That will sound somewhat abstract, so let us consider a concrete example. Suppose that we are considering an action which would increase people's motivation to pursue what they consider to be good consequences, at the expense of non-moral pursuits. However, suppose that we expect most of them to believe, on religious grounds, the claim that "it is right to bring about an abortion ban". Suppose that, while we accept the argument from Section 2.2.1 and admit that these people have somewhat truth tracking beliefs, we think people are biased: we judge them to be *very likely* to believe this claim even if it is wrong to ban abortion, and *almost certain* to believe it if it is right to ban abortion. Furthermore, suppose that we think banning abortion is much more likely to be wrong than right. Motivating these people to act on their moral beliefs would make them *much* more likely to try to ban—and, by the argument in Section 2.2.3, succeed at banning—abortion if it is right to bring about an abortion ban, and *somewhat* more likely to try to ban—and succeed at banning—abortion if it is wrong to bring about an abortion ban. So "it is right to bring about an abortion ban" judges our proposed action *very* positively while "it is wrong to bring about an abortion ban" judges our proposed action only *somewhat* negatively. Considered in isolation from our credence distribution, this fact, about how the two opposing theories judge the action, is a theory-neutral reason in the action's favor. However, when considered in combination with our relatively high credence in "it is wrong to bring about an abortion ban", it is also a theory-dependent reason against the action.

I said back in Section 1.3 that the stronger our theory-based reasons for or against a particular action are, the more likely they are to outweigh the theory-neutral reasons for or against it. This is such a case. It reminds us that we *do* have to pay attention to our theory-dependent reasons; it would be a mistake to assume that what we happen to have a theory-neutral reason to do will also be what we have *most* subjective reason to do.

I hope now to have convinced the reader that we really do have theory-neutral reasons to promote recognition, motivation, and success if we can. What remains is to show *that* we can. I shall discuss each of three connections individually, and suggest how one might set about trying to strengthen each of them.

2.3.1 – Promoting Recognition

The first way to strengthen the effect of goodness on which consequences occur is to enhance the "recognition" connection: to increase the overall likelihood of people making accurate moral judgments. I shall abbreviate this as "promoting recognition". Here is the statement of this theory-neutral reason:

> The *Theory-Neutral Reason to Promote Recognition* is our theory-neutral reason to perform actions which increase the likelihood that people—specifically, those who may be at least somewhat motivated to obey their moral value judgments and are likely to be at least somewhat successful at achieving their goals—will form accurate judgments about which consequences are morally good, especially when making morally significant comparisons.

Given that people are somewhat motivated by their moral judgments and are somewhat successful at doing what they are motivated to do, increasing the accuracy of people's moral judgments should increase the extent to which good consequence, whichever ones those are, will come about. So promoting recognition is one way to strengthen the connection between a consequence's goodness and its likelihood of occurrence.

Note that when I refer to the "accuracy" of people's moral judgments, I mean this in an objective sense. The goal given to us by this reason is *not* "cause others to believe the specific moral claims that we find most plausible". Certainly, if we knew what the true moral theory was, or even if we deemed some specific theories to be much more plausible than their opposites, this would give us a theory-based reason—at least if the theories in question were at-least-partly consequentialist—to teach people to believe *those specific theories*. But that is not what I am discussing here. In terms of theory-neutral reasoning, teaching any particular moral view is zero-sum: such teaching will be approved by that view but disapproved by that view's opposite. We have a theory-neutral reason to try to improve the *process* by which people, including ourselves, form their moral beliefs so that it will be more likely to lead them to *whichever* theory turns out to be true. It is a theory-neutral reason because—unlike merely replacing one specific belief with another—successfully improving people's reasoning process in this way will be approved by many at-least-partly consequentialist moral theories *and* their opposites.

Also note that the metric here is "accurate moral judgments about which outcomes are good", not "accurate moral judgments" simpliciter. This is for the reason already explained above: we are trying to perform actions which are right in and of themselves, not just actions which result in other right actions in the future. If an action makes agent-neutral goods more likely to be brought about, then by the definition of "good", that is at least a point in favor of its being a right action. On the other hand, if an action makes agent-relative moral duties more likely to be obeyed in the future, that is not necessarily a point in favor of its being a right action for *us* to perform *now*.

Speaking of the consequentialist nature of the theory-neutral reasons being discussed in this section, I use the word "promote" advisedly, as a word with consequentialist connotations. If an action has as its immediate effect a decrease in the accuracy of someone's moral value judgments, but has the remote effect of significantly increasing the accuracy of other people's moral value judgments, it *is* favored by the Theory-Neutral Reason to Promote Recognition. What matters is the effect, not how immediate the effect is.

The statement also includes the qualification "especially when making morally significant comparisons". What I mean here is that when evaluating the accuracy of a person's moral beliefs, we should focus on beliefs which are likely to be relevant to his decision-making, and in particular ones relevant to morally-important decisions. "*Ceteris paribus*, actions which prevent the extinction of trilobites are better than actions which do not" might be a true moral belief—biodiversity possibly has intrinsic value, and certainly has instrumental value for scientific research and artistic contemplation—but since trilobites have already been extinct for hundreds of millions of years, we are never faced with decisions in which one available option better prevents their extinction than another available option does. Similarly, "*ceteris paribus*, actions which prevent stubbed toes are slightly better than actions which do not" might also be a true moral belief, but the "slightly" makes it less important than, say, "*ceteris paribus*, actions which preserve many human lives are *much* better than actions which fail to preserve those lives". So when increasing the extent to which moral value judgments' popularity is correlated with their truth, we should focus not just on any old moral value judgments but on the most important ones—i.e. the ones which will actually influence people's decisions, and especially their most morally significant decisions.

What are the practical upshots of the Theory-Neutral Reason to Promote Recognition? Well, to start with, I mentioned in the Introduction the case of moral reflection. It seems to me that the point of engaging in moral reflection is the expectation that it will, on average, improve the accuracy of one's future moral beliefs. It is not *guaranteed* to do so, of course—some agents, as a consequence of reflection, undoubtedly end up exchanging accurate pre-theoretic intuitions for inaccurate post-theoretic judgments. However, so long as this is less likely than the alternative possibility that we will give up inaccurate pre-theoretic prejudices in favor of more accurate post-theoretic judgments, and so long as we *are* the kind of people who are morally motivated and not completely counterproductive in our goal-oriented behavior, it will have positive consequences, on average, for us to engage in reflection. If so, then the Theory-Neutral Reason to Promote Recognition will advocate moral reflection.

It is worth dwelling for a moment on the structure of this advocacy. For people to engage in moral reflection tends—provided, again, that reflection is on average not counterproductive, and that goodness really does matter in the way described in Section 2.2—to be instrumentally good. The theory-neutral reason does not specify what intrinsic good reflection is instrumental for achieving, but that is the beauty of theory-neutral reasons. If reflection helps one correctly identify and achieve intrinsically good outcomes, then it is instrumental toward intrinsically good outcomes, whichever outcomes they turn out to be. So we can advocate reflection without needing to know which outcomes are intrinsically good. Note, incidentally, that since I am identifying reflection *as an instrumental moral good*, not just as a morally right activity, anything that facilitates reflection is also advocated. We should engage in reflection ourselves, encourage others to reflect, save people time if they are likely to use that extra time for reflection, and so on. However, also note that I am merely claiming that reflection is *typically* morally *instrumental*, not that it is *always* morally right *all-things-considered*; if one has a theory-based reason to believe that engaging in reflection in a given situation, such as when trying to decide whether to flee a burning building, would be morally disastrous, this could easily outweigh the general theory-neutral reason to reflect. In fact, in the case of a burning building, the reason to reflect would not only be *outweighed* by the theory-based consideration that one's own life has intrinsic moral value, but would also be *undermined*, since if reflection during a fire causes one to die, then it will *not* increase one's likelihood of successfully pursuing moral goods in the future. In a choice between "flee a burning building" or "sit and ponder while the building burns around you", there is a theory-neutral reason to choose the former—doing so promotes recognition by preserving one's life and hence preserving one's ability to recognize moral goods—and no theory-neutral reason at all to choose the latter. In short, while there is *usually* a theory-neutral reason to engage in reflection, there is not *always* such a reason.

There may be other activities besides reflection which likewise improve one's moral judgment. Reading or listening to other people's moral arguments might be an example. So might experiencing a variety of lifestyles in hopes of perceiving moral value or disvalue in them. We could also try to inform people about known cognitive biases, and train them to resist those biases. All of these activities are advocated by the Theory-Neutral Reason to Promote Recognition. As in the case of reflection, we should engage in these activities ourselves, encourage other people to engage in them, and make it easier for other people to engage in them. Making it easier for people to learn about others' moral arguments might, for instance, involve funding a library or endowing a professorship at a university. Making it easier for people to have broad experiences from which they can derive moral judgments might involve setting up a free society, with mores encouraging experimentation.

This last idea is not new. John Stuart Mill gives a similar argument in *On Liberty* as one of his justifications for supporting individuality—he writes that allowing "experiments of living" will help demonstrate the value of different ways of life, thus facilitating moral progress.[36] By "valuable" he means "utility-promoting", but the argument does not depend upon that identity. Anyhow, this justification strategy for a free society is picked up on and elaborated by David Lloyd Thomas in his book *In Defence of Liberalism*, who gives it the name "experimental consequentialism":

> *Experimental consequentialism* is the view that society should be arranged so as to include the prerequisites for forming more reasonable [and therefore hopefully more accurate] views about what is intrinsically valuable.[37]

He contrasts this with "maximizing consequentialism", the idea that there is some already-identified value which we should try to maximize. While I agree with his attempt to separate his account from objective theories of value, I think that it is a mistake to frame it in terms of maximization versus non-maximization. It seems to me that the main reason anyone would have for following experimental consequentialism would be the theory-neutral reason sketched here. That is, we would be seeking information about what things are valuable not for the information's own sake, but as an instrumental step toward fulfilling those values.[38] But if so, then it *does* make sense to maximize: if two possible arrangements of society would both *permit* moral progress to occur, but one would allow moral progress to occur at a faster rate than its alternative would, we should—all else equal—choose the former. I say "all else equal" since of course if we have theory-based reasons, or even other theory-neutral reasons, in favor of choosing the latter, they might outweigh the Theory-Neutral Reason to Promote Recognition. Lloyd Thomas seems to be viewing experimental consequentialism as an isolated whole, whereas if I am right it is merely part of a larger framework: insofar as experimental consequentialism is a good idea, it is a good idea because it strengthens the link between which consequences are good and which ones occur; which, in turn, is a good idea because an action or policy which strengthens that link will have relatively high subjective rightness.

Notwithstanding this disagreement, I do think that the Theory-Neutral Reason to Promote Recognition *does* support attempts to create and maintain a free society, and I have more to say about that support. However, the Theory-Neutral Reason to Promote Recognition is not the only theory-neutral reason in favor of a free society, so I will be delaying further discussion of freedom until Section 3.4. For now I want to discuss ways in which the other two connections from Section 2.2 can be strengthened.

2.3.2 – Promoting Motivation

The second option for increasing the significance of goodness is to promote moral motivation. That is, we want to increase people's willingness to act in accord with moral beliefs—their own or others'—and their unwillingness to act in ways which significantly violate those beliefs. Recall from Section 2.2.2 that for purposes of this dissertation, actions count as "morally-motivated" even if they are not moved by the agent's moral beliefs so much as merely constrained by those beliefs: if an agent performs an action out of self-interest, but would not have performed it if he had believed it to be seriously immoral, his action counts as "morally-motivated". However, it *is* a requirement that the actions at least be constrained by moral beliefs, not just coincidentally in accord with moral beliefs—if the agent performs an action out of kindness or love, believes that morality approves of such actions, but would be performing it even if she believed that morality did not approve, her action does not count as "morally-motivated". In short, we want people's moral judgments, as often as possible, to play a decisive role in people's decision-making:

> *The Theory-Neutral Reason to Promote Motivation* is our theory-neutral reason to perform actions which increase the likelihood that people will be

> willing to pursue whatever consequences are judged—by themselves or others—to be better than the available alternatives, and unwilling to pursue whatever consequences are judged to be worse than some available alternative, especially when the moral difference between the options is judged to be large.

Given that people's moral value judgments are somewhat truth-tracking—more precisely, given that such judgments resemble the actual moral values more closely than people's whims or non-moral desires do—and given that people's attempts to bring about consequences are not normally counterproductive, increasing people's moral motivation will make them more likely to bring about good consequences.

As with the Theory-Neutral Reason to Promote Recognition, the focus is on people's consequentialist moral views, not their moral views generally. The reason is the same as before: theory-neutral reasons cannot assume the objective claim that we should try to cause people to act rightly in the future—whatever reasons we have to believe such a claim are theory-based reasons—but *can* assume the claim that we should try to cause people to bring about good outcomes in the future. They can assume the latter because it is analytically true: "good outcome" is defined as that which we should try to bring about, and one way to bring about an outcome is to cause someone else to do so.

The "especially when they judge the relative moral difference to be large" clause is also playing the same role as the "especially when making morally significant decisions" clause played during the discussion in Section 2.3.1. I do not advise that we simply increase people's bare likelihood of acting in accordance with beliefs about what values are good, but rather their weighted likelihood of acting in accordance with them—weighted by the moral significance they attach to each. If a person exerts extreme amounts of willpower to make tiny perceived improvements to the consequences of several mostly-morally-irrelevant decisions, and then is too tired to exert his will to make what would have been a major perceived improvement to the consequences of a single morally-very-important decision, he has erred; it would have been better if he had save his willpower for when it really counted.

Note, by the way, that the goal here is to actualize people's potentially-accurate moral beliefs, not just to create an accord between their beliefs and their actions. The Theory-Neutral Reason to Promote Motivation is in no way intended to support manipulating people into changing their moral beliefs to match what they would have done anyway, nor into construing decisions as morally more important in those cases in which they were already going to obey their moral judgment than in those cases in which they were already going to ignore it. Doing so would nearly guarantee that the people we were manipulating would *not* have truth-tracking moral value judgments; so after having manipulated them, their alleged "motivation" to obey their value judgments would be worthless.

In practice, how might we increase people's future willingness to pursue good consequences? This is, of course, an empirical psychological question. Figure out what values people in fact accept, and figure out what treatment encourages them to promote those values. I am not a psychologist, so not really qualified to say more here. However, I will offer some speculation. This *is* mere speculation; I have no real evidence for these ideas, so they should be tested before being implemented. I offer them mainly just to illustrate the sort of behaviors which *might* qualify as promoting the motivation element, to give an idea of what *kinds* of hypotheses to consider. They are the kind of actions that the Theory-Neutral Reason to Promote Motivation *might* advocate.

Here is my psychologically naive picture of how people make decisions: they weigh up all the various considerations for and against each available option—the extent to which they believe it to be consistent with their immediate self-interest, with their long-term self-interest, and with their personal projects, the extent to which they believe it to violate agent-relative duties, and the extent to which they expect it to result in morally good consequences—and then they choose the option which is most strongly favored, on net, by the aggregate of these considerations. If this is roughly right, then to increase people's chances of pursuing the consequences they think morally valuable, we need to do one of three things: increase the strength of their motivation to produce morally good consequences, decrease the strength of their other motivations, or align their other motivations with their judgments about moral goodness.

I confess to near cluelessness about how to strengthen the pull of moral goodness. The best idea I can offer is to try to draw people's attention to the similarities between agents: the fact that all of us have similar thoughts, feelings, desires, and so on. Hopefully this will cause them to feel intellectual pressure against treating the general good as less important than their own specific interests, or even their own agent-relative duties.[39] I take it that something like this is one motivation behind wanting ethnic diversity in public schools: the idea is that children of different backgrounds who interact frequently enough will come to see the deep similarities underlying their surface differences, and that this will somehow improve their characters.

As for decreasing the weight of considerations that pull against moral value judgments, that will depend on the details of those considerations. We should look at what distracts people from pursuing the good, and then try to remove those distractions. For example, people trapped in desperate poverty, struggling to survive, are not in a good position to direct effort toward moral goodness; the Theory-Neutral Reason to Promote Motivation favors raising them out of poverty if we can. Likewise we should try to cure addicts of their addictions. We should also probably investigate the old idea that willpower—the ability to resist the pull of one's immediate self-interests in favor of other goals—is trainable.[40] Perhaps if children are frequently put in situations in which they have a strong incentive to resist desires for immediate self-gratification, or even in situations in which they will be unable to satisfy such desires, they may end up being better equipped to resist such desires when those desires are in conflict with moral beliefs.

Last comes the possibility of trying to realign people's non-moral motivations to stop conflicting with their value judgments. We could draw their attention to the warm, fuzzy feeling one gets from doing a good deed, so that acting morally will seem more consistent with their own happiness. We could try to convince them that actions with morally bad results will frequently be detected and punished by authority figures—elders, the law, god, whoever—while actions with morally good results, even if apparently violating rules, will not be punished. We could even go a step further and teach that all rules, even moral rules, are ultimately justified by the consequences of following them; if we can argue people into consequentialism, they will be less likely to knowingly pursue bad consequences out of concern for deontological duties.

That is probably more than enough armchair psychology. If my suggestions—e.g. encouraging people to take impartial viewpoints, training their willpower or breaking addictions, trying to increase the perceived alignment of their self-interests with agent-neutral morality, etc.—can be shown by empirical psychology to increase people's tendency to do what they themselves believe to have morally good consequences, then the Theory-Neutral Reason to Promote Motivation will endorse those suggestions. If not, it will endorse whatever methods *do* increase people's tendency to do what they believe to have morally good consequences. I shall now leave the empirical science to empirical scientists and return to more philosophical issues.

2.3.3 – Promoting Success

The final way to increase the significance of consequences' goodness for their likelihood of occurring is to increase people's success at doing what they try to do.

> The *Theory-Neutral Reason to Promote Success* is our theory-neutral reason to perform actions which strengthen the connection between the consequences people attempt to bring about—at least when motivated by truth-tracking moral judgments—and which consequences occur.

The idea is that a goal which is motivated by a truth-tracking moral judgment is more likely, all else equal, to be morally good than morally bad, so making it more likely to come about is a morally right action. Increasing the correlation between people's goals and what happens makes such goals more likely to come about. As I did in Section 2.2.3, I am neglecting the possibility of an idealistic goal which can be approached but not achieved: count that as partial fulfillment.

It is worth noting here that the purpose here is to strengthen the *overall* connection between what people are trying to achieve and what happens; we are using people's potentially-value-tracking goals as an indication of which consequences are good. We are not, however, placing an intrinsic value on giving those people the *feeling* of success. For example, suppose that a given person, call him John, has found an injured bird and wants this bird to recover and return to the wild. Suppose an experienced wild bird rehabilitator, call her Mary, is in a position to help, and can do so in one of two ways. She can give John enough advice and supplies for him to successfully rehabilitate the bird himself, or she can persuade him that the bird is dead and then secretly rehabilitate it without his knowledge. As far as the Theory-Neutral Reason to Promote Success is concerned, both of these options equally count as fulfilling John's desire for this bird to recover—even though the latter gives John much less role in the fulfillment, and never makes him aware of the fulfillment having taken place.

The usual caveats are also present. If we have theory-based reasons to suspect that a given goal—even though it may have been formed in a value-tracking manner—would not, in fact, be morally good to fulfill, then these reasons will weigh against the Theory-Neutral Reason to Promote Success. If they are sufficiently strong, helping to fulfill the goal will not be the subjectively right action; instead we should work toward the consequences, if any, which are supported by the theory-based reasoning. Theory-neutral reasons are not all-things-considered reasons; it is subjectively right to follow them only when they are not outweighed by other considerations.

Another caveat is that, just as the Theory-Neutral Reason to Promote Motivation was concerned with causing a person's actions to fit his moral beliefs, but *not* with causing his moral beliefs to fit his actions, the Theory-Neutral Reason to Promote Success is concerned with causing the outcome to fit his goals, *not* with causing his goals to fit the outcome. If we manipulate a person into adopting whatever goals are easiest to fulfill, those goals would not be tracking moral value.

The perceptive reader will have the following worry about whether the Theory-Neutral Reason to Promote Success truly counts as theory-neutral. Learning that someone with value-tracking goals has a particular goal *does* give us a reason to fulfill that goal, but it does so by giving us new information about the goal in question: it increases our credence in the moral theories which hold that goal to be intrinsically morally good, and decreases our credence in the moral theories which hold that goal to be intrinsically morally bad. This sort of thing was supposed to be ruled out by the part of the definition of theory-neutral reasons given in Section 1.1, which said that theory-neutral reasons raise an action's subjective rightness in a way *other* than by raising our credence in some theories above our credence in their opposites.

This is a concern about how to classify our reasons, not about what reasons we have. Nevertheless, it is right as far as it goes. If we know a person's goals, our reason to promote success at those goals is not technically a theory-neutral reason. However, suppose that we do *not* know the content of a person's goals; all we know is that the goals are—or might be—value-tracking, and that a given action is likely to help the person fulfill those goals. In that case, our reason to perform that action is a theory-neutral reason. If utility is valuable, the goal in question—since it is value-tracking—is relatively likely to involve utility, so the action in question—helping the person fulfill that goal—is relatively likely to bring about utility. If equality is valuable, the goal in question is relatively likely to involve equality, so the action in question is relatively likely to bring about equality. And so on. So if we want to be technical, the Theory-Neutral Reason to Promote Success applies only to promoting the success of goals whose content is not known to us; there is a parallel Non-Theory-Neutral Reason to Promote Success which applies to the goals whose content *is* known to us.

Once again, it is crucial to understand what is and is not being argued. I am not making a theory-based argument that the objective theory "it is right to do help people with value-tracking goals succeed at doing whatever they are trying to do" is more likely to be right than "it is wrong to help people with value-tracking goals succeed at doing whatever they are trying to do". What I am arguing is that a great many *other* objective theories and their opposites, theories whose axioms do not even *mention* "people with value-tracking goals", will nevertheless instrumentally approve of helping people with value-tracking goals succeed at doing whatever they are trying to do.

Now the reader may have a related question. If we do not know what a person is trying to do, how are we supposed to help him succeed at it? To answer this, it is important to distinguish *proximate* goals from *ultimate* goals. Proximate goals are consequences which a person pursues simply because they are instrumental to his ultimate goals. Ultimate goals are consequences which a person pursues for their own sakes, e.g. because he believes them to be morally good, or because he feels that his life will have gone better if they occur. Everything I have said above about the value of promoting success has been about ultimate goals. We have a theory-based reason to help people's value-tracking ultimate goals succeed when we know what those ultimate goals are, and a theory-neutral reason to help people's value-tracking ultimate goals succeed when we do now know what those ultimate goals are.

It frequently happens that we *do* know a person's proximate goals despite not knowing his ultimate goals. For example, if we see someone trying to nail two sheets of plywood together, or if he tells us this is what he is trying to do, we can reasonably infer that he thinks this will be instrumental to *something*, but we may not be able to guess what purpose he intends the construction to serve. Nevertheless, under the Theory-Neutral Reason to Promote Success, we should try to help him nail the two sheets together. If he is trying to pound the nail with an ill-suited stick, we might offer him a hammer to use instead; if he asks us "which of these nails do you think is the right length to secure these boards together?" we might give him our best advice about the question; and so on. Even though we do not know what his ultimate goal *is*, we know that it can be served by nailing together the plywood, so helping with that is a way to help with the goal.

How to help with any given proximate goal depends heavily on the features of the proximate goal and of the general situation, so I have little more to say about it here. A more interesting case is when we do not even have any special information about what proximate goals a person has. This would be true of a stranger whom we have never met. It is also true of future people, including our future selves—specifically, future people who we can expect to have value-tracking goals *later*, but who may not have those goals *yet*. I think that in cases of strangers or future people, it is still possible to help the people in question despite not knowing anything about what specifically they want, or will want, to do.

What makes this possible is John Rawls's insight that there are "goods [which] normally have a use whatever a person's rational plan of life". He calls these "primary goods".

> A *primary good* is something which will normally be useful to a person regardless of what his goals are.

Rawls's gives a list of "chief primary goods", mentioning "rights, liberties, and opportunities", "income and wealth", "self-respect", "health and vigor", and "intelligence and imagination".[41] They are goods which can be proximate to *any* ultimate goal which is chosen—or at least to most such possible goals.

I add the qualifier "at least to most" because primary goods, as defined here, are not *always* useful for a person. Instead, their structure is rather like that of theory-neutral reasons themselves: some of a stranger's possible goals come in pairs which cancel each other out, while most possible goals would be furthered by his possession of primary goods even though their opposites would too. Here is an example to show what I mean. Suppose that we can give a small child a vaccination that makes him immune to some crippling disease, but that the vaccination will only work if given before he is old enough to make the decision for himself. Our intent in giving it to him would be to increase his future health, a primary good. It is possible that in the future he will have a goal of having a body completely free of unnatural chemicals, or will have a goal of wanting to live a life of challenge and hardship; if one of these is his goal, he may resent having been vaccinated.[42] On the other hand, he might also have the opposite goals of having a body as heavily-upgraded as possible, or of living a life of health and ease—goals which would be strongly advanced by vaccination. If these were his only four possible goals and were equally likely for all we knew, vaccination would be equally likely to be a hindrance as a benefit. However, they are not the only possible goals he might have. Instead of caring about bodily integrity or personal challenges, he might enter politics and try to advance a conservative agenda, or enter politics and try to advance a progressive agenda; either way, health will be an asset. He might form the goal of living in the wilderness isolated from human influences, or of living in a city and being connected to the pulse of civilization; either way, health will be an asset. And so on. Some pairs of possible goals are such that improved health would hinder one, if that one is chosen, while advancing the other if it is chosen instead; but some—I suspect most—pairs are such that improved health would advance *whichever* is chosen. Hence the claim that improved health is more likely to be help than hindrance even for a future person about whose ultimate goals we know nothing, that improved health is a primary good. A similar picture will hold for other primary goods; they have the capacity to be used in service of most possible goals, even ones which are mutually exclusive.

Since I am claiming that we have a theory-neutral reason to increase people's access to primary goods, I wish to make a few specific comments about particular items from the list. First, it should be noted that the advocacy of health, vigor, intelligence, imagination, and self-respect—in short, the physical and mental traits which help a person accomplish things—goes beyond advocacy of mere disease prevention. Given an opportunity to increase the capacities of a human organism beyond those to which we have been accustomed in the past, for example by discovering new techniques of education, or even via some kind of genetic enhancement, we have reason to do so. If creative geniuses are better able to fulfill their goals than regular people are, let us produce creative geniuses. Let us do so, that is, provided that we exercise due caution against unforeseen side-effects—obviously we should not lightly abandon methods that worked well in the past. I shall return to the issue of human genetic engineering in Section 4.1.3; for now I only note that there might be theory-neutral reasons to engage in it.

The other item I want to highlight is "income and wealth", a short phrase which glosses over a very large set of goods. One aspect of wealth is for people to have enough food, water, shelter, and medical care to maintain their physical health and vigor. Another aspect is for people to have access to the materials they need to carry out their work: carpenters need tools, nails, and wood; authors need ink and paper; chefs need ingredients and heat; and so on. A third aspect of wealth is intellectual; people need access to information and training for how to achieve their chosen goals. Then there are social resources, such as being able to identify other people with proximate goals similar to one's own, and coordinate with them. Note also that I say *access to*, not *possession of*, with respect to all of these different forms of wealth. What is important is that, for example, a person who wants to write be able to acquire ink and paper, not that he already have it; and people in a society with a hundred books in a public lending library might well have greater total *access* to books than they would in a society with a thousand books in private collections. The key is to make sure that the wealth is present in society, and that the structure of society is such that individuals can draw on that wealth. Trying to increase total *access* to wealth is *not* the same as trying to increase total holdings; when the two come apart, it is the former that is favored by the Theory-Neutral Reason to Promote Success.

Notwithstanding the issue of arranging society so as to ensure that people have access to its wealth, it is nevertheless true that, *on average*, it will tend to be better for there to be more materials goods, more information, etc., in existence. The more wealth there is, the easier it will tend to be for any given individual to purchase, borrow, or otherwise access the portion of that wealth which his proximate goals require. So the Theory-Neutral Reason to Promote Success supports actions which increase the total wealth of society: all else equal, the more useful materials and information we produce, and the fewer we use up, the better. In concrete terms: industry and science tend to be morally desirable, while waste tends to be morally undesirable.

There are a handful of exceptions to the theory-neutral reason to increase total access to materials and information. For instance, some material goods or technical skills have as their main function the interference with other people's activities—I have in mind especially weaponry. It is unlikely that giving everyone in the world access to nuclear weapons, thereby allowing a few individuals in a small amount of time to destroy what millions of individuals have spent their lives producing, would increase people's overall ability to achieve their goals. I will discuss the problem of conflicting goals in later chapters; for now, simply note that weapons-related advances—e.g. inventing a more accurate sniper rifle—are not so obviously endorsed by the Theory-Neutral Reason to Promote Success as non-weapons-related advances—e.g. inventing a more efficient agricultural technique—are. Aside from the danger of making it easier for people to interfere with each other's plans, other possible exceptions are goods or knowledge which are socially destabilizing or psychologically overwhelming. For example, the grief and distraction caused by hearing about a tragedy about which an agent is unable to do anything might make it more difficult for him to focus on projects which *are* within his power, making that particular bit of information worse than useless.[43] In short, some common sense is necessary here—increasing wealth, health, knowledge, and so on are *instrumental* goals toward increasing people's ability to steer the world toward whatever outcome they think best, not ends in themselves.

My last note about primary goods concerns rights and liberty. They are indeed primary goods; however, *power over others* is also a primary good in the sense defined above; although not necessarily essential, being able to order other people about will tend to be useful for a wide range of possible projects. The Theory-Neutral Reason to Promote Success endorses actions which increase people's freedom from arbitrary rules, habits, and superstitions—actions which increase people's freedom without producing a corresponding decrease in anyone else's power—but it would be too quick for me to say here that it also endorses actions which increase people's freedom from one another's domination. Freeing previously-dominated people is good for their own goals but bad for their former dominators' goals; so there is a sense in which it does not increase total resources so much as just transfer a resource—the freed people's labor—from one group's control to another. I do think that rights and liberty are a good idea with respect to overall goal-fulfillment, but the explanation for that is complicated and will have to wait for Section 3.4's discussion of how to allocate goods which many different people want to use on their projects.

Incidentally, one action endorsed by the Theory-Neutral Reason to Promote Success is something which many of us already perform as a matter of general prudence, namely trying to increase our *own* future ability to achieve whatever projects we may *later* decide are desirable. Consider the proverbs "do not burn your bridges" and "save for a rainy day". "Do not burn your bridges" means that we should avoid restricting our future options, even ones which we do not presently intend to use. "Save for a rainy day" means that we should stockpile resources for dealing with presently-unforeseen problems. In both cases, the idea is that even though we do not know why our future selves might want a given option or a given resource, we trust that our future selves, if they do want it, will have a good reason for wanting it. The proverbs are good prudential advice insofar as the good reason might turn out to be prudential; however, insofar as the reason might instead turn out to be moral, the proverbs are also good moral advice.[44] Prudence, however, only tells us to work toward our *own* success, whereas the Theory-Neutral Reason to Promote Success tells us to increase *everyone*'s success—this is because self-interest is agent-relative whereas moral goodness, as defined above, is not.

To summarize this section: given some plausible claims about people's abilities and habits, I have shown that acts which increase people's ability to recognize morally good outcomes, their motivation to pursue those outcomes, or their success in that pursuit will tend to be approved by the vast majority of at-least-partly consequentialist theories, including those which are opposites of one another. So we have a theory-neutral reason to perform such acts. Therefore, unless the theory-neutral reason is counterbalanced by sufficiently strong theory-based reasons against performing such acts, they will be the acts which are morally right in the subjective sense described in Section 1.1.1. That is, although not necessarily the right actions according to the objective truth, they are right with respect to our credence distributions across moral theories and physical hypotheses, and so are the best we can do given our uncertainty about those matters.

2.4 – Other Theory-Neutral Reasons

The above three theory-neutral reasons—the Theory-Neutral Reasons to Promote Recognition, Motivation, and Success—are the ones I believe to be most important. However, they are not exhaustive. In this section, I will briefly discuss two possible variations: first, a Theory-Neutral Reason to Imitate Others, and second, the possibility of agent-relativizing theory-neutral reasons. These are of interest primarily just as examples of theory-neutral reasons that fall outside my main framework. However, they might also be of interest to readers who are concerned about my dependence on the possibility of an at-least-partly consequentialist moral theory being true. The Theory-Neutral Reason to Imitate Others would apply even to someone who was certain that all moral rules were deontological; e.g. taking the form "no one should ever perform A, even in circumstances where failing to perform A has the consequence that many more A-ings will occur over the long run". Agent-relativized theory-neutral reasons would apply even to someone who was certain that all moral rules were agent-relative, e.g. taking the form "members of group X should try to make consequence C occur; non-members of group X should try to prevent consequence C from occurring".

2.4.1 – Imitating Others

Suppose that an agent sees an impoverished panhandler sitting next to the sidewalk, and notices that many other passersby are tossing coins into his pan. Should the agent also toss a coin into the pan? Of course the agent should take into account whatever theory-based reasons support or oppose the action, and should also take into account the theory-neutral reason to help the panhandler fulfill whatever morally-motivated goals he might have. But what should the agent make of the actions of the other passersby? Do they have some consequentialist moral goal in mind, a goal that might also be fulfilled by giving to the panhandler? Suppose that it seems unlikely. Any consequentialist moral goal they might have—reducing poverty and inequality, say—would be better served in some other way, such as by giving the money to a homeless shelter or soup kitchen. They are acting on non-moral reasons, or at least on non-consequentialist reasons. So as far as the Theory-Neutral Reason to Promote Success is concerned, their decision to give money to the panhandler is irrelevant to whether the agent should give money to the panhandler.

However, I think that it is not entirely irrelevant. By an argument analogous to the one in Section 2.2.1, if the passersby think they have a non-consequentialist moral duty to give money to the panhandler, then this is a reason to think that they are right. Maybe the true moral theory includes some deontological rule like "when an agent is faced with someone less well-off than himself, he should give a token of his concern—regardless of how the overall consequences of this action compare with those of possible alternative actions". If so, then the rule that the other passersby have recognized might well apply to the *agent* as well.

For the most part, of course, this should simply raise the agent's credences in moral theories which approve of the action others are performing. However, just as we can help someone fulfill a goal without knowing what goal is at issue, it may also be possible to imitate someone's obedience to a purely non-consequentialist rule without knowing what rule is at issue. For example, maybe the passersby are tossing coins into the panhandler's pan in obedience to the non-consequentialist rule "do not ignore the plight of the badly off"; the coins represent a token acknowledgment of the panhandler's plight. But maybe the passersby are tossing coins into the panhandler's pan in obedience to the non-consequentialist rule "do not be distracted by the plight of the badly off"; they are giving coins because they happen to be aware that as a fact of human psychology, they are less likely to dwell upon the panhandler's situation later if they feel like they offered some token help and can imagine everyone else doing likewise. In short, seeing passersby perform actions of type A may raise the agent's conditional credence that A obeys rule R if R is true, but also raise the agent's conditional credence that A obeys the opposite of R if the opposite of R is true. If so, then the agent could have a theory-neutral reason to imitate their action and toss a coin in the pan.

> The *Theory-Neutral Reason to Imitate Others* is our reason to try to perform the same types of actions as other people are performing, insofar as those actions might be motivated by moral judgments—even non-consequentialist ones.

This is, of course, a fairly weak reason. Given that the agent does not know why people are taking a given action, the agent may have trouble identifying the relevant category of action. For example, in the panhandler case, is it just "toss money into the pan", or is it something more specific like "toss a penny—but definitely not a quarter or dollar—into the pan, with the intention of acknowledging the panhandler's presence but not incentivizing panhandling", or something else entirely like "if you are planning to enter a building with a metal detector and a long line to pass through it, remove spare change from your wallet in advance and discard it in such a way that it does not end up as litter or in a landfill"?

Also, the Theory-Neutral Reason to Promote Recognition would probably recommend that the agent simply ask the other passersby whether they are acting on a moral rule, and if so what it is and what reasons they have for accepting it. That way the agent could make an informed choice—not to mention participate in the overall dialogue and thereby help others to make better choices as well. Blind imitation, in short, is a waste of one's mental resources and so will commonly be overruled by other theory-neutral reasons. Except for people who are certain that purely non-consequentialist moral theories are the way to go, it is unlikely to play a large role in the overall structure of theory-neutral ethics. However, weak and usually-overruled or not, the Theory-Neutral Reason to Imitate Others *does* qualify as a theory-neutral reason, and as one that is at least somewhat distinct from the ones discussed in Section 2.3.

2.4.2 – Relativization to "People Like Us"

It seems very possible that moral duties, or at least some of them, turn out to be agent-relative, but in a way that systematically assigns identical duties to agents that belong to the same groups. For example, perhaps everyone has a patriotic moral reason to promote the flourishing of *his own* nation, beyond whatever moral reasons he has to promote human flourishing generally. Perhaps membership in a particular religion can place special moral obligations on an agent, to obey that religion's commandments and pursue that religion's goals, which simply do not apply to non-members.[45] And so on for membership in other categories: species, extended families, social clubs, maybe "race" if it can be defined rigorously enough to ground moral rules, etc.

Some moral duties could also be *time*-relative. For example, perhaps everyone has a reason to increase aggregate well-being. But perhaps the appropriate aggregation scheme is one that discounts far-future well-being in favor of more immediate well-being. So someone living in the twenty-fifth century should worry about increasing the well-being of people in the twenty-sixth century; but someone living in the twentieth century should not concern himself with twenty-sixth century well-being. For another example, perhaps everyone should try to increase equality, where the relevant measure of equality includes cross-temporal comparisons. So whether one may perform an action which reduces the well-being of people living in the far future might depend on how those people's average well-being compares with the average well-being of present people. But perhaps the only relevant comparisons are with people who do exist or will exist, not with ones who used to exist: how present people's average well-being compares with the average well-being of people in the distant past might *not* be relevant to whether one may perform an action which reduces the well-being of present people. So someone living in the twentieth century might be able to partially justify actions which reduce the well-being of people in the twenty-sixth century by appealing to the fact that those people, even after the reduction, will still be extremely well-off by twentieth-century standards; but someone living in the twenty-fifth century might not be able to make this appeal.

Finally, and far more plausible than group-relative and time-relative duties, there might be moral duties than are individual-relative. For example, an individual might have a moral duty to pursue some state of affairs—e.g. one in which *his own* promises are kept, or in which *his own* debts are paid, or in which *his own* dependents are protected. Other people might have an equivalent obligation regarding *their* promises, *their* debts, or *their* dependents, but they will not have an obligation, or at least not as strong an obligation, to help the individual fulfill *his* duties.

In light of these possibilities, we could add "especially people similar to us" clauses to all of the above theory-neutral reasons. We have a theory-neutral reason to promote recognition of moral truths, *especially recognition by people similar to us*. We have a theory-neutral reason to promote motivation to act morally, *especially motivation of people similar to us*. We have a theory-neutral reason to promote successful goal achievement, *especially success of people similar to us*. And, for whatever it may be worth, we have a theory-neutral reason to imitate others, *especially others who are similar to us*.

How significant these "especially people similar to us" clauses are depends on how likely it is that morality is agent-relative, and to what extent. Personally I doubt that much of morality is group-relative; but that is a theory-based judgment, and so is not something I should defend here even if I had more to defend it with than bare intuition. It also depends, again, on whether we can identify the relevant categories. Who counts as more similar to an agent in morally-relevant respects: a present-day neighbor who attends the same church as the agent does, sends her children to the same school as the agent does, and so on; or the agent's own future self fifteen years down the line? How about someone from a different country and a different culture but with a similar profession, similar personality, and similar lifestyle to the agent's, versus someone from the agent's own country and culture who has a very different profession, personality, and lifestyle?

Even if we somehow discovered that morality was entirely group-relative, we would still have some theory-neutral reason to promote *everyone's* recognition of moral duties, motivation to follow them, and success at following them—and to imitate *everyone's* non-consequentialist-but-still-morally-motivated actions—simply on the grounds that *anyone* might, for all we know, belong to the same morally-relevant group of people that we do.

CHAPTER THREE – NEUTRAL POLICY

In this chapter, I will argue that theory-neutral reasons favor acting in accordance with—not *accepting as objectively true*, but acting in accordance with—something resembling a utilitarian moral code and a version of a liberal political code. The comparison with utilitarianism and liberalism will highlight several philosophically interesting features of the application of theory-neutral reasons. It will also shift the burden of proof regarding the practical question of whether to obey the theories with which theory-neutral reasons accord—we will no longer need to justify decisions to obey those theories, but will instead need to justify decisions to *disobey* them.

Before I begin, however, I need to acknowledge a difficulty: what we should do in any given situation depends on the balance of reasons in that situation—*all* reasons, both theory-neutral and theory-based. This balance will vary from situation to situation, so there is little I can say of much generality while avoiding taking a stand on theory-based considerations. However, there is a related question which I *can* productively discuss here: the question, not of what we should do, but rather of what procedures we should follow when deciding what to do. When making decisions *now* about how *future* decisions will be made—e.g. when deciding what habits to cultivate in myself and encourage in others, what norms to popularize, what legislation to support, etc.—most of the details of the situations in which those future decisions will take place are simply unknown, so perforce must be neglected. So I *can* say some general things about how to make this sort of decision.

A second advantage of focusing on this indirect question is that it allows me to avoid the technicality mentioned in Sections 2.3.3 and 2.4.1 about the blurriness of whether a given consideration counts as a theory-neutral reason or a theory-based reason. Recall the issue: our reason to increase a stranger's effectiveness, or to imitate his actions, is a theory-neutral reason, since it is based on our conditional credence that whatever unknown values or rules the stranger is basing his action on are relatively likely to be correct; whereas our reason to increase a friend's effectiveness, or imitate his actions, is a theory-supporting reason, since it is based on the extra credence—already informed by our familiarity with his views, and so not conditional—which our trust in his judgment causes us to place in the *particular* values or rules upon which we know him to be basing his action. In both cases the reason is derived from our trust in the other agent's judgment; but its classification is different in the two cases. If I tried to discuss what theory-neutral reasons we have in particular situations, I would therefore be forced to say confusing things such as "we have a theory-neutral reason to help strangers but not to help friends". I think it is much less confusing to take a step back and say "we have a theory-neutral reason to adopt a policy of helping both strangers and friends"—which I *can* say, since if such a policy were adopted sufficiently far in advance, we would not know who our friends were going to be nor what values they were going to have.

Anyhow, this chapter will focus on that indirect question of how we should *now* plan to make *future* decisions. Call what is at issue here the *Neutral Policy*:

> The *Neutral Policy* is the set of habits, rules of thumb, decision procedures, norms, laws, etc., which we have most theory-neutral reason to try to get people, including ourselves, to follow in the future: i.e. the ones which best encourage recognition of true moral values, motivation to pursue those values, and success at that pursuit.

This chapter will examine the content of the Neutral Policy, and compare and contrast it with familiar utilitarian and liberal policies of the sort advocated by John Stuart Mill.

If we could, before encountering any theory-based reasons, have chosen a policy for dealing with those reasons, we should have committed to the Neutral Policy. Not, of course, because we place more credence in theories which explicitly say "commit to the Neutral Policy" than in theories which say the opposite, but because we place some credence in opposing pairs of theories whose axioms do not explicitly mention the Neutral Policy at all, but which nevertheless tend to favor acting—including making commitments—on the basis of theory-neutral reasons.

Of course, the fact that it *used* to be subjectively right for us to commit to the Neutral Policy does not mean that it is still subjectively right; that will depend on how our theory-based reasons weigh into the picture. Also, effort spent trying to institute the Neutral Policy—within our society's laws, within individuals' decision procedures, etc.—is effort not spent on theory-based projects or on other theory-neutral projects; so whether and to what extent we should exert such effort will depend on how our various reasons stack up. The Neutral Policy is not the whole of morality, nor even the whole of what we have theory-neutral reason to do. Nevertheless, I think it is worth discussing: the question of which norms to advocate is, after all, fairly central to moral philosophy.

3.1 – Goal Fulfillment

We have a Theory-Neutral Reason to Promote Success; i.e. to maximize the extent to which the things people try to bring about actually come about, at least insofar as such maximization does not conflict with other theory-neutral reasons. So the Neutral Policy will include an instruction to be helpful: i.e. to help people fulfill their goals. For now I will focus on that instruction; I will return to the Theory-Neutral Reasons to Promote Recognition and Motivation in Section 3.2.

Incidentally, when I say that we should help people fulfill their goals, that most definitely includes ourselves. One should not sacrifice all hope of fulfilling one's *own* goals, or the goals one may adopt in the future, simply to slightly improve someone *else's* chances of fulfilling *his* goals.

The Theory-Neutral Reason to Promote Success can be compared to the moral theory of utilitarianism, which I take to be the theory that it is objectively morally right to maximize well-being, and making the comparison will be the major focus of this section. Note that, in making this comparison, I do not claim that this should increase our credence in the versions of utilitarianism which bear the most similarity to the Neutral Policy. The Neutral Policy has a completely different justification from that of any objective moral theory, since it is grounded in theory-neutral reasons, and since those reasons involve what we *subjectively* ought to do given our moral uncertainty rather than what we *objectively* ought to do if only we knew it. However, I think it is still useful to ask about the extent to which "act *as though* you were a utilitarian", "be beneficent", or "maximize people's well-being" is an accurate summary of our theory-neutral reasons.

Different kinds of utilitarians interpret "well-being" in very different ways. They can be clustered into roughly three families.[46] First are hedonistic utilitarians, who hold that what is good for a person is to have positive mental states such as pleasure, happiness, or self-esteem, and not to have negative mental states such as suffering, boredom, or anxiety. Hedonistic utilitarians can disagree about *which* mental states are morally significant, and how to weigh those states against one another, but they agree that mental states are what matters. Second are preference utilitarians, who hold that what is good for a person is for his preferences and desires, or some subset thereof, to be satisfied or fulfilled. Of course, in general we feel happy when our preferences are satisfied and unhappy when they are frustrated, and, for that matter, in general our preferences tend to feature a desire to have positive mental states and not to have negative ones. However sometimes preference satisfaction and positive mental states come apart; sometimes we want things which will not make us happy. For example, a person who wants to know whether his friends respect him might, if the answer is "no", be happier not knowing; a hedonistic utilitarian might think it is good for this person to be deceived, or to be distracted from the question, while a preference utilitarian would say that if the person truly prefers to know the answer, it is in his interests to find out that answer, notwithstanding how it will make him *feel*. Third are ideal utilitarians, who might be inclined to give some list of traits—e.g. self-realization, living an objectively moral life, having true beliefs, having friends, etc.—as the components of well-being, and say that it is good for a person to have such traits regardless of how having them makes him feel and regardless of whether he wants to have them.

To the extent that the Theory-Neutral Reason to Promote Success has anything to do with promoting well-being, it has to do with the "preference fulfillment" family of interpretations. It tells us to help bring about whatever consequences people have adopted as goals. It is not concerned with whether bringing about those consequences will make people happy, nor with whether bringing about those consequences will help those people develop other allegedly-ideal traits. True, once we add in the Theory-Neutral Reason to Promote Recognition and the Theory-Neutral Reason to Promote Motivation, it will start to be concerned with people's epistemic state and their motivations, so we might try to compare it with a version of ideal utilitarianism which says "maximize the extent to which people have true beliefs, good character, *and* fulfilled preferences". But I shall argue in Section 3.2 that there is a better and more precise way to incorporate the other two theory-neutral reasons within the Neutral Policy. Until then my discussion will focus just on the Theory-Neutral Reason to Promote Success, and the idea that we should just maximize people's level of goal-fulfillment.

Of course, "help people succeed at their goals" is not the same as helping people get what they want, since a person might choose as a goal something which he does not want for himself and does not regard as being in his interests, if he does regard it as impersonally morally good or as in someone else's interests. Preference utilitarianism would be in danger of circularity if allowed such goals to be taken into consideration. Imagine that everyone accepted such a version of preference utilitarianism as his moral theory, and that everyone was so disciplined and moral that he never pursued any goals except the one which preference utilitarianism gave him. So everyone would have the goal of living in a world in which everyone's goals were fulfilled. Would these goals count as fulfilled, under that version of preference utilitarianism? "Yes" and "no" are both internally consistent answers, which suggests that something has gone horribly wrong. So preference utilitarians should probably focus on fulfilling *non-moral* desires, and maybe even only fulfilling *self-interested* desires, not on fulfilling moral goals.[47]

In contrast, the Neutral Policy is not subject to this kind of problem, and does not require any sort of restriction to some subset of a person's goals. Unlike utilitarianism, it does not purport to be an objective theory of how morally good any given situation—such as the situation in which everyone's only goal was goal-fulfillment—would be; instead it purports to be a subjective strategy for making ourselves more likely to bring about a relatively good situation rather than a relatively bad one. So it has the leeway to offer advice which is useless in some strange situations, as long as the advice is generally good. And since the Neutral Policy does not purport to be an objective theory of morality, the situation in which people are following the Neutral Policy and the Neutral Policy only—rather than balancing it against other moral beliefs—definitely counts as a strange one.

Notwithstanding this distinction between self-interested preference fulfillment and moral goal fulfillment, preference utilitarianism and the Neutral Policy do have many similarities in the kinds of actions they recommend. Insofar as one way to advance a stranger's morally-motivated goals is to give him primary goods which will enable him to fulfill *all* of his goals, both morally-motivated and self-interested, the two will have significant overlap. So it is worth looking at further issues within preference utilitarianism and seeing whether they apply here.

Within the broad family of preference utilitarian interpretations of well-being, there is room for disagreement about what it means to fulfill a preference. In particular, what if the person does not know that his preference has been fulfilled? For example, consider the following case:

> *The Case of the Starving Artist*: Arthur was an artist who all his life wanted, and strove to cause, people to appreciate his artwork. He was not particularly trying to achieve this appreciation *during his lifetime*, however. For example, when he knew that he was dying, he chose to spend his final hours putting the finishing touches on a new work, despite knowing that nobody would see it until after his death; if he had been aiming for appreciation *in his lifetime* rather than appreciation *simpliciter*, he would have instead spent those final hours throwing an exhibition of

> his existing works, to get as many people as possible to see them before his death, in hopes that someone would appreciate them. In any case, his works were never appreciated during his lifetime, and he died in obscurity believing that he had probably failed. There turns out not to be any sort of afterlife; so at the instant of his death, he, his preferences, and his goals ceased to have present existence. After this death, people discover Arthur's works, recognize their merit, and begin to appreciate them. Do these events count as retroactively fulfilling his preferences and goals?

Different preference utilitarians will have different reactions to this case. Is it truly in an agent's interests that his preferences be fulfilled *simpliciter*, or must he also *know* that they are fulfilled or even *enjoy* the fact that they are fulfilled? Relatedly: is it truly in an agent's interests for his preferences to be fulfilled *eventually*, or must they be fulfilled *during his life* or even *while he still holds them*? Preference utilitarians will schism over these questions, with some considering the above scenario to be good for Arthur and others considering it to be bad for him. It will depend on how far they are willing to distance themselves from common-sense hedonism, and also on their attitudes toward metaphysical statements such as "the past and future exist".

The Neutral Policy, however, should *not* schism. To extent that followers of the Neutral Policy try to fulfill Arthur's goal of having people appreciate his work, we are not doing it for *his* sake. We are not thinking that his possession of the goal *causes* appreciation to be morally good; rather, we are thinking that his possession of the goal is *evidence* that appreciation is morally good. That is, his goal might, for all we know, be at least partly based on a truth-tracking moral judgment—something along the lines of "all else equal, it is right to perform actions which cause people to have positive aesthetic experiences", perhaps, or perhaps some other moral consideration. However, whatever he was thinking, we can infer from his behavior—from the fact that his goal was clearly "that people appreciate my artwork" and not "that people appreciate my artwork during my lifetime"—that he thought the value of such appreciation would not vanish with his death. So we should proceed on the assumption that he may be right, that by contributing to his goal we will *also* be acting in a worthwhile manner. For my purposes:

> A person's goal counts as *fulfilled* if the state of affairs involved in the goal comes about, even if it comes about at a time when the person no longer has the goal. The Neutral Policy advocates the fulfillment of goals in this sense of "fulfillment".

Of course, it may often be the case that a person's goal will, if carefully specified, be that some event occur *during his lifetime* or that it occur *while he still wants it to occur*. If the event occurs but *not* during his lifetime or *not* while he still wants it, this will of course *not* count as fulfilling his goal. To say otherwise would be to ignore part of the content of his goal. We are indifferent to whether the goal is fulfilled while it exists *only* if the content of the goal makes no reference to the goal's own existence.

This can be expanded into a more general point about the importance of correctly describing a person's goal. Situations such as "S is trying to make it the case that X occurs", "S is trying to make it the case that X occurs while S still wants it to occur", "S is trying to make it the case that S directly causes X to occur", and so on, can be difficult to tell apart from one another. However, they are *not* equivalent; to the extent that we *can* tell them apart, they warrant different behavior under the Neutral Policy.

In the context of correcting for mistakes, it is worth returning to the distinction, from Section 2.3.3, between a person's ultimate goals, and the proximate goals he is pursuing only for the sake of fulfilling those ultimate goals. Often we will not know what the ultimate goals are, and so the only way to try to fulfill them will be by trying to fulfill the proximate ones. However, when we *do* know what the ultimate goals are, the proximate goals cease to be of any concern to us at all. For example, consider the following scenario:

> *The Case of Busy Beth:* Beth has a goal of having a flower garden in her front yard, where many people—including herself—will be able to enjoy looking at the flowers. She also has a proximate goal of having some spare time, so that she could use that spare time to plant the flower garden. One day you have the opportunity to help Beth in one of two ways: you can plant the flower garden for her, or you can do some chores for her so that she will have just enough free time to be able to plant the garden herself. Both options satisfy her ultimate goal of having a flower garden, but the latter also satisfies her instrumental goal to have more free time.

The Neutral Policy should be indifferent here, despite the fact that doing other chores for Beth would help with her proximate goals—assuming that Beth's goals are exactly as described, there is no reason to treat it as important that *she* be the one to plant the garden.

Incidentally, it is worth noting that goals formed on the basis of theory-neutral reasons are inherently instrumental. We want to promote moral recognition, motivation, and success not for their own sake but because they are expected to result in good consequences. So as people learn about theory-neutral reasons and form goals on the basis of them, we do not gain *new* reasons to promote those goals. If Hal has the goal of helping Fanny be motivated by her moral beliefs, and his motivation for pursuing this goal is entirely based on theory-neutral considerations, then while the Theory-Neutral Reason to Promote Motivation still tells us to join in, for the sake of motivating Fanny, the Neutral Policy does not suddenly give us a second reason to join in, for the sake of helping fulfill Hal's goal. This is another reason why the circularity worry mentioned above is not a threat to the Neutral Policy: the goals it favors fulfilling in others are not the kinds of goals it tells us to adopt for ourselves.

Related to the focus on ultimate goals, and to the general fact that we are concerned with goals only insofar as they might be value-tracking: we need not concern ourselves with *mistaken* goals. If we suspect that a person adopted a goal only because he was not thinking clearly or was misinformed in some important way—this will especially tend to happen with proximate goals—then we are probably justified in ignoring that goal. Indeed, we quite possibly should instead try to fulfill the goal he *would* have if he had *not* made the mistake, if we can identify it. For example, suppose that a given person, call him Alfred, is trying to assassinate Brian, Charles, and David. Suppose that the only salient feature which Brian, Charles, and David have in common with one another is that Alfred believes they were responsible for a terror bombing which took place several years earlier. Suppose that we know that in fact the responsible parties were Brian, Charles, and *Donald*; David is innocent. Insofar as we trust Alfred's *moral* judgment—which appears to be that some to-us-unspecified moral theory is true and that punishing the bombers will be judged positively by that theory—despite being convinced that his factual judgment about who committed the bombing is wrong, we might well have a theory-neutral reason to *prevent* Alfred from fulfilling his goal of killing David, at least until we can acquaint him with the evidence against Donald and see how that changes his proximate goals.

I mentioned early the possibility of restricting preference utilitarianism's domain of concern to self-interest preferences: preferences a person has about how his own life should go. Followers of the Neutral Policy may also be interested in whether a person's goals have this feature, but will react to it in exactly the opposite way: far from ignoring goals which do not have this feature, we can *discount* goals which *do* have it. The grounds for such discounting is that goals like that that they are relatively likely to be influenced by non-moral motivations such as self-interest, and so relatively *unlikely* to be value-tracking. In particular we should discount goals whose fulfillment would produce positive mental states in the person pursuing them, since these are the sort of goals which most people's conceptions of self-interest will lead them to pursue while not morally motivated. This makes the Neutral Policy *very* non-hedonistic; it will tend to favor sacrificing people's hedonistic interests for the sake of non-hedonistic ones. Consider the following scenario:

> *The Case of the Food Bars:* You are planning to ship food aid, in the form of high-calorie nutrition bars, to a starving region. There is an option regarding flavor: they can be delicious chocolate bars or equally-nutritious but relatively-flavorless soy bars. The bars will be equally nutritious either way. Which option should you choose?

I think the Neutral Policy favors the soy bars here. The reason is that people who need calories to carry out the projects they believe to be moral will find the soy neither more nor less helpful than the chocolate, since the two flavors are stipulated to be similarly nutritious. However, people who are not being motivated by judgments of moral goodness will be more prone to misuse the chocolate, for example by consuming it even when it would be healthier for them to avoid the extra calories.[48] The lesson here is that the Neutral Policy is not just more comparable to preference utilitarianism than to hedonistic utilitarianism, but is in fact *very* non-hedonistic; it places no weight at all on people's pleasure *per se*, no matter how much they prefer to experience it. As a result it can advocate actively avoiding the fulfillment of hedonic desires if it is worried that they are overruling moral judgments.

This last example should not be carried too far; while the Neutral Policy is non-hedonistic, it is not so anti-hedonistic as to actively favor the frustration of hedonic desires. It is concerned with fulfilling moral goals, so seeks to prevent hedonically-motivated goals from playing a significant role in what happens; but it does not seek to actively frustrate hedonically-motivated goals, since this *would* give them a role. If the choice in the above example had been between flavorless soy beans or disgusting lima beans, our reasons would favor the soy just as strongly as in the choice between flavorless soy and delicious chocolate; just as we would not want people to pursue the food on hedonistic grounds when it would not benefit their moral projects, we also would not want them to *avoid* the food on hedonistic grounds when it *would* benefit their moral projects. So at worst the Neutral Policy favors a stance of calmness and neutrality toward hedonic goods, not deliberate seeking out of pain or suffering.[49] Furthermore, in general our hedonic desires are paired with primary goods—it *feels good* to acquire primary goods like health, safety, social status, and so on; and *freedom from constant pain* may well be a primary good in itself since it can be difficult to pursue any project when too severely distracted by pain. Lastly, it is of course true that, at least at present, moral theories including "all else equal, it is right to cause others pleasure and wrong to cause them pain" enjoy far higher popularity than their opposites; and causing unnecessary pain would frustrate the morally-motivated goals of people who subscribe to such theories. So on the whole, I suspect that obeying the Neutral Policy would result in the creation of far more pleasure than pain, notwithstanding its non-hedonistic nature.

I would be remiss not to point out that the Neutral Policy does not precisely advocate giving weight to goals in proportion to their likelihood of being motivated by moral beliefs—it instead advocates giving weight to goals in proportion to their likelihood of being motivated by *moral beliefs formed via recognition of the moral truth*.[50] If some people's moral beliefs were more likely than average to have been formed by a truth-tracking process, we would want to give more weight to those people's goals. I shall return to this point in Section 3.3's discussion of how to weigh one person's morally-motivated goals against another's; for now I just note it and move on.

Another feature relevant to whether a goal is value-tracking is its position in time. I defined "fulfillment" above in a way which allows the fulfillment of goals which no longer exist; but *should* we fulfill such goals? At the very least, we might want to know what happened to them. If the goal went away due to the person changing his mind about what to pursue, this might be evidence that it was based on a mistake which he subsequently caught. If it went away due to the person making some fatal error of practical judgment and dying, this could cast doubts on his general reasoning ability and hence on the accuracy of his moral judgments. However, if the reason why the goal in question is non-present is something more normal—e.g. the person whose goal it is died of old age, or is yet to be born—then there is less reason to discount it. Dying does not retroactively cause a person's moral judgment to worsen; absent new information, whatever credence we assign to "so-and-so's goals are value-tracking" while he is alive should continue to be the credence we assign to "so-and-so's goals *were* value-tracking" after he is dead. So there is at least some theory-neutral reason to pay attention to a person's last will and testament, deathbed requests, and so on; the goals involved are worth fulfilling despite being in the past. If anything, a person's goals involving what happens after his death are very clearly *not* motivated by hedonism—assuming he believes that he will no longer be capable of having experiences after he dies—so are relatively likely to be motivated by moral judgment, and so should receive some *extra* weight. That is not to say that they are *certain* to be morally-motivated; a person can have egocentric, non-moral desires for things which do not require his own survival, such as a desire that his enemies perish. But *many* types of non-moral desires will be irrelevant to a person's goals about what happens after his death, which is why I say that those goals are *relatively* likely to be morally-motivated.

However, while the Neutral Policy does not suddenly discount goals after the person with those goals dies, I suspect that it will *gradually* discount goals over time—it will not place the same weight on goals of people from the *distant* past as it does on present people. The reason is intellectual progress. It appears to me that humanity grows wiser over time, accumulating potentially-morally-relevant knowledge and ideas. I might be wrong—I might just be living in an upswing of a more cyclical process—but if I am not wrong, then a person living early in history is less likely to have value-tracking goals than a person living late in history, all else equal. Note that this only holds within a given civilization's continuity: if we found a thousands-year-old message from extraterrestrials, or if history does turn out to be cyclical and we found a message from an advanced "Atlantis" civilization from thousands of years ago, there would be no particular reason to discount whatever goals it advocates. Note also that, far from privileging *present* goals, the consideration about position in history favors *future* goals. We should be more concerned with fulfilling our own morally-motivated goals than with fulfilling the morally-motivated goals of people from ancient history; but we should be more concerned with fulfilling our distant descendants' morally-motivated goals—assuming we can forecast what actions on our part will prove useful to them—than with our own. I will discuss this point in more depth in Section 4.2.

More can be said about how to try to focus our efforts at goal fulfillment on value-tracking goals at the expense of non-value-tracking goals, but I think it more profitable to move on to a new question: that of *whose* goals matter. Once again, preference utilitarians have room to disagree about this issue. Some of them will think that we should include non-human animals within the sphere of moral concern; others will think that we should limit our maximization of well-being to *human* well-being. There are also borderline cases such as mentally disabled humans, or science-fiction cases such as conscious computer simulations of humans.

The Neutral Policy, however, will again take a clear and univocal stance on the question of whose goals matter. We are not satisfying goals for the sake of the beings that hold them; we are satisfying goals for the sake of the true moral values which they might be reflecting. So *if* a given being's goals are potentially reflecting true moral values, they should be taken into account. Who or what the being in question happens to be is irrelevant.

In practice, of course, not very many beings will have goals of the right sort. I suspect that any being with an intelligence significantly lower than that of a normal adult human—whether it is a non-human animal, or a mentally-disabled human, or an infant human—will be incapable of making potentially-truth-tracking moral value judgments. If I am right about this, then the Neutral Policy ignores the goals of all such beings—not because of who they are, but simply because of what kind of goals they have. Similarly, if there exist beings which are capable of making moral value judgments but not of basing their goals on their judgments—perhaps some psychopaths qualify; and I can imagine artificial intelligences which had this feature due to having their goals hardwired into them—their goals would also be ignored.

I should make a few qualifications to these comments. One involves domesticated "partnership" animals such as dogs and horses. I take it that such animals, although they probably do not make moral judgments of their own, can be trained to have goals which reflect *their trainer's* moral judgments. If we see a St. Bernard rescue dog trying unsuccessfully to pass through a gate, and we do not know *why* it wants to get through the gate, we might well have a theory-neutral reason to help it do so, on the grounds that its ultimate goal might well be reflective of true moral values. Of course, in some sense what we are doing is trying to satisfy *the trainer's* preferences by helping the dog, so this is not necessarily a counterexample to "only worry about the preferences of beings capable of making truth-tracking moral value judgments".

Another qualification would involve animals which are basing their goals not on training but on their instincts, but whose goals somehow reflect moral judgments. One might hold such a view if one believed that the world was, or might have been, designed by a morally good God who shaped animals' instincts with an eye toward producing good consequences. Atheistic examples can also be given. For example, a sufficiently lengthy breeding program might produce dogs which, with no need for any training at all, have goals reflecting their breeders' moral judgments; we would want to offer assistance to those dogs too, just as we offered assistance to the dog who had been trained to pursue morally-good consequences. For a subtle variation: suppose that when humans first migrate to a given region of the globe, they kill off all of the animal species except the ones they deem to be doing morally useful things. If so then we might conceivably have reason to help even wild animals do whatever they are trying to do, for the sake of the judgments of the long-dead ancestors who decided to spare those animal species.

How about a dumb animal species which has never even interacted with beings capable of making moral judgments? Could its members ever have value-tracking goals? I am inclined to think not. The animal could of course display altruism, or feel "moral" emotions like empathy, but the evolutionary explanation for these traits would be something like "there is evolutionary pressure to treat possible kinsmen altruistically" or "there is evolutionary pressure to treat altruistically individuals that might reciprocate this treatment". These are not value-tracking mechanisms, but rather mechanisms that would tend to produce such traits regardless of whether such traits had morally good or morally bad effects. So insofar as we have a moral reason to help such animals fulfill their goals, it will be a theory-based reason—one based on a judgment that helping the animals will happen to serve valuable ends. It will not be a theory-neutral reason based on a judgment that the goals in question are value-tracking.

I have tried to argue in this section that the Theory-Neutral Reason to Promote Success favors a "Neutral Policy" of goal fulfillment. In many cases this Neutral Policy will favor the same values as the objective moral theory of preference utilitarianism does, and I have used this accordance to explore the Neutral Policy in more detail. It does not merely overlap with preference utilitarianism generally; it overlaps specifically with a version which is interested in fulfilling preferences even after those preferences have ceased to exist, which ignores preferences which are based on mistakes, and which restricts the sphere of moral concern to beings capable of moral reasoning. Of course, the two are not identical. Aside from having completely different grounding—the Neutral

Policy is based on subjective considerations of trying to bring about good consequences no matter which consequences turn out to be objectively good, whereas utilitarianism is based on the judgment that consequences involving preference satisfaction are the ones which are objectively good—the Neutral Policy focuses only on moral goals, with complete disdain for non-moral desires.

3.2 – Creating Goals

I now turn from the question of which goals should be fulfilled to a different issue: what to do when it is within our power to alter which goals exist. After all, the number of moral agents, and what morally-motivated goals they have, is not fixed. Children can be brought into the world and become new moral agents, with morally-motivated goals which are to some extent a function of how those children are raised; existing moral agents can suffer death or severe brain damage and thereby cease to be moral agents.[51] Meanwhile, existing moral agents' views change over time, to some extent as a function of what experiences they have. As a result, many of our most important decisions change not only which goals are fulfilled but which goals will exist in the future, and do so in at least partly-foreseeable ways. I said above that future goals are relevant to the Neutral Policy, perhaps even more than present or past ones are; but our ability to manipulate the number and content of future goals can create complications. This section will address those complications, and in the process show how to assimilate not just the Theory-Neutral Reason to Promote Success, but also the Theory-Neutral Reason to Promote Recognition and the Theory-Neutral Reason to Promote Motivation, under the policy of goal fulfillment.

The two most common utilitarian responses to the question of when to create or destroy moral subjects are total utilitarianism and average utilitarianism, favoring the maximization of total utility and of average utility respectively.[52] To see how they could come apart, suppose we are considering two possible actions, A and B, whose effects are equivalent in most morally-relevant respects, but that choosing A will bring an additional person into existence who would never exist if we choose B. Total utilitarianism favors A if the additional person's utility would be "positive" and B if the additional person's utility would be "negative". So total utilitarians need to specify a "zero point"—they must specify how well-off a person must be to count as neither increasing nor decreasing the total, which means evaluating people's utility on an absolute scale rather than in solely comparative terms. This creates a range of possibilities, each of which results in a distinct version of total utilitarianism. For example, "a subject's utility is positive if she is glad that the universe exists, and negative if she wishes that it had never come into existence in the first place"[53] and "a subject's utility is positive if, on the whole, more of her preferences—weighted by the significance she attaches to them—are satisfied than frustrated, and negative if more are frustrated than satisfied", both of which I find at least a little bit plausible, are definitely distinct views. Average utilitarianism, on the other hand, favors A if the additional person's utility would be above-average and favors B if the additional person's utility would be below-average. There is no need for average utilitarians to define an absolute scale of utility or a zero point; all they need to be able to do is compare one person's utility with another's.

Note that these are not alternatives to the kinds of utilitarianism discussed in Section 3.1. Instead, they are variations along a separate dimension: there can be total hedonistic utilitarians, total preference utilitarians, average hedonistic utilitarians, and average preference utilitarians.

These approaches may be the most familiar ones in the realm of utilitarian theories about objective morality, but *neither* is a suitable analog for the Neutral Policy. To see this, consider the following case:

> *The Case of Ambitious Amber and Boring Bob:* Ambitious Amber and Boring Bob have different personalities, which has led them to form different moral judgments and hence different morally-motivated goals. Amber is the kind of person who is constantly looking for opportunities for improvement, and rarely spends much time thinking about what could go wrong. She adopts morally-motivated goals such as "ending world hunger"; ones which are very unlikely to be fulfilled no matter what she does, but which she supports nevertheless—if she raises the probability of success from 1% to 1.01%, she will judge her time to have been well spent. Bob, on the other hand, is the kind of person who appreciates what he already has, and is highly risk-averse. He adopts morally-motivated goals such as "preventing the collapse of civilization"; ones which are very likely to be fulfilled no matter what he does, but which he supports nevertheless—if he raises the probability of success from 99% to 99.01%, he will judge his time to have been well spent. We have every reason to believe that both Amber and Bob will, if given the chance, continue forming new goals in the future along the same pattern. However, one day an unspecified disaster threatens the lives of Amber and Bob, and we can save only one. Whom should we save?

If we were concerned with maximizing total goal fulfillment *or* average goal fulfillment, we would save Bob rather than Amber, since saving him will probably result in new morally-motivated goals coming into existence which are likely to be fulfilled, whereas saving her will probably result in new morally-motivated goals coming into existence which are then likely to be frustrated. However, choosing Bob on these grounds is *not* consistent with the Theory-Neutral Reason to Promote Success. As I noted in Section 2.3.3, the goal of the Theory-Neutral Reason to Promote Success is not to manipulate people's goals so that they will match what happens, but only to manipulate what happens so that it will match people's goals. If we judged that Amber's goals were less important than Bob's, or that Amber was making less of a difference than Bob was, those might be reasons to save Bob rather than Amber; but the mere fact that Amber's goals are less likely to be fulfilled than Bob's is not a reason to save Bob rather than Amber.

A better—for my purposes—alternative could be called "existing preference utilitarianism". Instead of holding that we should maximize total preference fulfillment or average preference fulfillment, existing preference utilitarianism holds that we should maximize the fulfillment of preferences which already exist, or whose existence is already assured, at the time of our decision. So when deciding whether to bring a given preference into existence, existing preference utilitarianism says that we should ignore that preference itself and instead look only at preferences whose existence is *not* contingent on the decision at hand. For example, when deciding whether to produce a child, existing preference utilitarianism would tell us to look at the effect the child's life is likely to have on others, but not at whether the child itself is likely to have satisfied preferences. Incidentally, there is no need to specify whether we are concerned with maximizing the total fulfillment of the existing preferences or the average fulfillment of the existing preferences, since the size of the set of existing preferences at the time of a given decision cannot be affected by that decision, and so the total fulfillment is directly proportional to the average fulfillment.

A similar approach should be taken by the Neutral Policy. Instead of trying to bring morally-motivated goals into existence which are likely to be fulfilled, we should try to make it the case that many of the morally-motivated goals *which already exist* get fulfilled and few of the morally-motivated goals *which already exist* are left unfulfilled.

However, while this approach may be the right thing to say about the Theory-Neutral Reason to Promote Success, it is not a good approach for the three Theory-Neutral Reasons when taken as a whole. Consider the following scenario:

> *The Case of Blowing Up the World*: You have the opportunity to destroy the world, abruptly extinguishing all life and permanently putting an end to the formation of new morally-motivated goals. You weigh up everyone's existing possibly-morally-motivated goals. Many goals, such as promotion of happiness or advancement of knowledge, would be frustrated by the cataclysm; but a few very important ones, such as prevention of severe suffering, would be fulfilled by it. On balance, when all the goals have been weighed up, it turns out that blowing up the world and refraining from blowing it up would be equally satisfying of existing goals. Should you blow up the world?

Even setting aside theory-based reasons not to blow up the world—e.g. the fact that "it is morally very wrong to kill billions of people" is much more plausible than "it is morally very *right* to kill billions of people"—and considering the problem solely from the perspective of theory-neutrality, I think it would be a mistake to flip a coin here.

Consider: there may be morally-significant values about which we currently have no clue. Compare how likely those currently-undiscovered values are to be fulfilled in three different scenarios: first, if we blow up the world; second, if we do not blow up the world, but nobody ever discovers those values; third, if we do not blow up the world, and we eventually discover those values. If we are genuinely clueless about the contents of those undiscovered values, then we should estimate their likelihood of fulfillment in the first two scenarios as being equal. But from the discussion in Section 2.2, we should estimate their likelihood of fulfillment in the third scenario to be higher than their likelihood of fulfillment in the second. Since refraining from blowing up the world has some chance of resulting in the third scenario rather than the second, the currently-undiscovered values are *more* likely to be fulfilled if we do not blow up the world than if we do. Since the currently-discovered ones, as represented by existing goals, were supposed to be ambivalent between blowing up the world or not, the balance of *all* values can be expected to *oppose* blowing up the world.

This line of reasoning against blowing up the world is an application of the Theory-Neutral Reason to Promote Recognition, not of the Theory-Neutral Reason to Promote Success—the logic, at its heart, is that more recognition of morally-significant values will take place if there are people surviving to recognize those values. The Neutral Policy needs to be able to capture this sort of reasoning.

We can do this by admitting *potential* goals into consideration, without regard to whether they are actual goals.

> A goal counts as *potential* with respect to a given decision if it is a goal that somebody *could* come to have as a result of the decision. The Neutral Policy tells us to try to fulfill—that is, try to bring about the states of

> affairs which are or would have been favored by—anyone's potential goals to the extent that those goals are or would have been value-tracking, regardless of whether those goals become actual.

This view—unlike a version which focused only on existing goals—gives the right answer in the "Blowing Up the World" case: it has us take into account the morally-motivated goals which would come about if we refrain from destroying humanity, and notice that the states of affairs favored by those goals are more likely to come about if we do refrain from destroying humanity than if we destroy it. However, it also—unlike a version which naively totaled or averaged actual goals—gives the right answer in the "Ambitious Amber and Boring Bob" case: if Amber's and Bob's potential goals matter regardless of whether they become actual, then we can only focus on trying to make the states of affairs those goals would favor as likely to occur as possible; those states' background likelihood becomes irrelevant. Preventing Amber from forming her ambitious goals would be like killing the messenger: it would not eliminate the bad news of the goals' difficulty.

Even in absence of these thought experiments, it should not be surprising that the Neutral Policy regards potential but non-actual goals as significant. I said back in Section 3.1 that a reliable judgment does not suddenly become less reliable when the person who formed it dies; it is still the case that the judgment *was* reliable, even though the judgment is no longer present. I now offer this corollary: a reliable judgment does not suddenly become less reliable when it is precluded from being formed; it is still true that the judgment *would have been* reliable, even though the judgment is no longer nomologically possible. Put in more familiar terms: if we somehow knew that a hypothetical impartial observer would morally disapprove of an action, this is a reason to suspect that the action is immoral. And it is not as though I were claiming that the fulfillment or frustration of merely potential goals were *intrinsically* significant, as the analogous version of utilitarianism—"*potential* preference utilitarianism"?—would be committed to doing. That would be rather strange, at least absent belief in modal realism; it would be bad enough for counterfactual preferences to have actual moral relevance, but for them to be *as* relevant as actual preferences would be a stretch indeed. But anyhow, I am not claiming that here. My argument is simply that when we fulfill or frustrate merely potential goals, we may also thereby be advancing or impeding the *actual* values which those potential goals could have tracked. It this *actual* advancement or impediment which is actually morally significant; the potential goals are just being used as part of our subjective strategy for pursuing undiscovered values.

Of course, we cannot know much about the content of reliable judgments which are precluded from being formed; after all, if we do not know what the true moral values are, but *do* know exactly what the contents of some counterfactual judgment would have been, then clearly the judgment in question was not reliably value-tracking and there is no reason to try to fulfill it. In the end, I suspect that the *only* practical difference between "try to fulfill all potential goals insofar as they may be value-tracking" and "try to fulfill existing goals insofar as they may be value-tracking" is that the former—given the argument from Section 2.2.3—favors the *creation* of new value-tracking goals even when this creation is neutral with respect to existing goals. Other than that, "we ought to try to fulfill merely potential value-tracking goals" will give advice every bit as tautological and useless as "we ought to try to bring about morally good consequences, i.e. those consequences which it is right to bring about, whatever they are".

That said, I still prefer "we ought to try to fulfill all potential goals insofar as they may be value-tracking" to "we ought to try to fulfill actual goals insofar as they may be value-tracking, and also try to bring about the adoption of value-tracking goals" as a statement of the Neutral Policy. Phrasing it as two instructions, as the latter formulation does, would leave would-be policy-followers wondering how to weigh fulfillment of value-tracking goals against adoption of value-tracking goals. Phrasing it as one instruction does not. We should bring about a potential value-tracking goal if and only if that promotes the fulfillment of value-tracking goals generally, potential or actual.

Incidentally, this captures *all three* Theory-Neutral Reasons from Section 2.3. The reasons given for promoting recognition and motivation were that they contributed to the formation of value-tracking goals. This completes my comparison of the Neutral Policy with utilitarianism: the Neutral Policy amounts to "try to fulfill people's goals", while preference utilitarianism says the similar "try to fulfill people's preferences", but we have seen that there are important disanalogies between the two. Section 3.1 argued that the Neutral Policy is concerned with goals only to the extent that those goals might be value-tracking, so tends to be much less hedonistic, and much more focused only on beings with advanced mental capacities, than preference utilitarianism is; and the current section has argued that the Neutral Policy is concerned with fulfilling even goals which are merely potential rather than actual, which would be strange in a theory of objective rightness like utilitarianism.

What remains to be done is to address something I have not yet explicitly discussed: how the Neutral Policy would have us make interpersonal comparisons, weighing one person's morally-motivated goals against another's. That will be the task of the next section.

3.3 – Democracy

So far I have argued that if we can increase the extent to which some people's goals are fulfilled, *without decreasing the extent to which others' goals are fulfilled*, then the Neutral Policy supports doing so. However, normally it is not that simple; normally there are tradeoffs. As noted earlier, one person's freedom comes at the expense of another person's power—and both freedom and power are useful for almost any project. Likewise, allowing one person to use a given material resource will typically mean that other people will not have the opportunity to use it. Information can be shared freely, but its discovery or creation bears costs; producing information that benefits one project may come at the expense of producing information that benefits another. We have just seen that in situations of overpopulation, bringing one person into existence makes it harder to advance existing values—if the new person shares the existing values, then this represents a loss for everyone, but if he does not, then this will represent a tradeoff between benefiting his values but setting back existing ones.

I take it that the utilitarian approach to tradeoffs is "everybody to count for one, nobody for more than one".[54] That is, in the case of a tradeoff between one group's values and another's, utilitarianism will, all else equal, tell us to support whichever side has the greatest number of people in it.[55] The Neutral Policy will agree, since it only has us trying to help people fulfill their goals insofar as those goals may have been motivated by *accurate* moral beliefs. On the assumption that each person's moral beliefs are weakly truth-tracking—that is, each person is thought to be more likely to believe any given objective moral theory if it is true than if it is false, but is still not guaranteed to believe it—it follows that a belief shared by many people is more likely to be true, all else equal, than a value shared by fewer.[56] So one way we might resolve tradeoffs is democratically: siding with what the majority wants in each case.

However, all else is rarely equal. Even utilitarians do not look only at the number of people on each side of a given issue; they will also look at the importance each person places on the issue. The Neutral Policy should do likewise. A small group which is *confident* that its values would be advanced, and would be advanced significantly, from a decision in their favor quite possibly—provided, of course, that its confidence is justified by evidence and argument, rather than just being a symptom of different group norms of confidence-ascription—ought to be able to overrule a large group which is less confident that its values would be advanced by a decision in its favor, or which expects a much smaller advancement.

How do we weigh the number of people on each side of an issue against their average degree of interest in having the issue resolved in their favor? Utilitarianism would give a simple answer, directly weighing the two considerations against one another. Given a choice between providing a large or certain benefit to a small group of people or a benefit half as large or half as certain to a group of twice as many people, utilitarianism will be indifferent. On the other hand, what the Neutral Policy will say is another story. It should not automatically give twice as much weight to the opinion of a majority twice as large as the minority opposition. Instead, it should base whatever extra weight it gives to the majority on the relative prior likelihood that the majority would be right and the minority wrong. This is *not* a simple function of relative group size; it will be much more complicated than that.

A few numerical examples will demonstrate this point. Suppose that on a given issue, each of three people has a 60% chance of favoring the morally best option and a

40% chance of favoring the morally inferior option, and suppose also that which option any one of them favors is independent of which option any other favors. Then the probability of getting a 2-to-1 split in favor of the best option is about 43%, while the probability of getting a 2-to-1 split in favor of the inferior option is about 29%—somewhat more than half as high as the other. But now modify the case slightly, keeping everything the same except having 30 people instead of 3. Now the probability of a 20-to-10 split in favor of the best option is about 12%, while the probability of a 20-to-10 split in favor of the inferior option is about 0.2%—much less than half as high as the other. For a third case, imagine that there are two inferior options, so each person has a 60% chance of favoring the best option and a 20% chance of favoring each bad option. Under those conditions, the probability of a 2-to-1 split in favor of the best option is still about 43%, but the probability of a 2-to-1 split in favor of a particular inferior option drops to about 19%—also less than half as high as the other, although much closer than the 20-to-10 case. So factors we should take into account when deciding how convincing a majority opinion is include not just the relative size of the majority as compared to the minority, but also the absolute size of the population and the number of options being considered.

Matters get even more complicated when we relax the independence constraint, and accept that people frequently get their moral beliefs from *one another* rather than arriving at those beliefs as individuals. For example, suppose that a religious leader, one of his followers, and a moral philosopher are debating a given issue. Suppose that the religious leader and the philosopher each have an independent 60% chance of favoring the best option and a 40% chance of favoring the inferior option. Suppose that the follower will favor whatever option the leader tells him to favor—so, since the leader has a 60% chance of favoring the best option and a 40% chance of favoring the inferior option, so does the follower. This is just like the first case from before except that one of the beliefs is no longer being formed independently. But now the probability of a 2-to-1 split in favor of the best option is 24% and the probability of a 2-to-1 split in favor of the inferior option is also 24%, so the fact that a 2-to-1 majority favors an option in *this* scenario is not at all persuasive as a reason to accept it. More generally, when comparing the numbers on different sides of any given issue, the Neutral Policy should ignore people who base their goals unquestioningly on what others have told them.

Of course, there is room for a middle ground between complete independence and complete interdependence. Suppose the follower puts *much* faith in his leader's positions, but does not adopt them completely unthinkingly; he does examine them a little bit, and has some chance of opposing his leader if he thinks he sees a good reason to do so. In this case, while his opinion should not be counted for as much as an independent opinion, it should not be ignored entirely. Or suppose an even less credulous follower, one who respects the leader enough to listen to the leader's rationale, but then makes up his own mind. In that scenario, we should not weight the two as though they were completely independent thinkers, especially in the case where there are multiple independent options—it is easier for one person to make a mistake and for another to fail to catch that mistake when listening to the first's argument, than for two people each to make similar mistakes—but we should definitely give them more collective weight than we would give to a single person.

On the other hand, even if two people reach their conclusions without communicating with one another, they may still not count as precisely independent. For example, suppose that two economists, with similar methods and similar preconceptions, but who are not actually in communication with one another, set out to construct models to forecast the outcome of a given policy. Suppose that there are a wide range of possible outcomes, but nevertheless the two economists end up making very similar forecasts. This might well be a sign that they both reasoned well and got accurate answers, or it might be a sign that their similar methods led them to make similar mistakes. Contrast this with a case in which an economist bases his forecast on abstract game-theoretical models while a historian, who in addition to belonging to a different academic discipline also comes from a different culture, bases her forecast on generalization from historical events. Here we can no longer explain away agreement between them by appealing to similar mistakes—unless perhaps the mistake was something general to all mankind such as "allowing wishful thinking to skew the reasoning"—but would instead have to postulate that the two made different mistakes during their different procedures which nevertheless coincidentally skewed the forecasts in similar ways. The alternative explanation for their agreement—namely, that both methods resulted in accurate forecasts—becomes increasingly appealing.

To the extent that it is feasible to do so, the Neutral Policy should take all of these gradations—from blind acceptance to careful review to similar methods to nearly complete independence—into account when tallying up the relative numbers on each side of a tradeoff. It will probably severely discount the goals of children whose moral views match those of their parents or other nearby authority figure. Adults whose moral views appear to have been received from somewhere else—religious leaders or texts, talking heads on television or radio, etc.—should also be discounted relative to independent thinkers, although not *completely* discounted since at least they have had *some* opportunity, poorly exercised though it may have been, to evaluate the views or at least decide *who* to place their trust in. Thoughtful people who nevertheless stem from a single monolithic culture should be discounted slightly relative to thoughtful people who spring from isolated cultures—linguistic divisions may be a good measure of this, both as a cause and a symptom of isolation, although we should also take into account merely ethnic or geographic divisions, and even divisions between academic disciplines. And so on.

People who have inherited their views from others rather than arriving at them independently are not the only ones the Neutral Policy will discount when weighing desires against one another. We should also discount those whose goals are relatively unlikely to be motivated by recognized moral truths. I already said earlier in this chapter that we should not concern ourselves with the goals of beings which do not have the concept of morality—I speculated that this category includes infants and non-human animals. The reader may have suspected that this line, between beings which can have morally-motivated preferences and beings which cannot, is somewhat fuzzy. It did not matter in the earlier discussion, since the Neutral Policy favors the fulfillment of any preferences that even *might* be morally-motivated, but it does matter once we start weighing preferences against each other and making tradeoffs.

Which people are more likely than others to have and be motivated by accurate moral beliefs? Answering this question mostly depends on meta-ethical questions about how we can discover the moral truth. If the best way to get accurate moral beliefs is by engaging in moral reflection and argument, then extra weight should be given to the people who have devoted the most time and effort to creating and examining moral arguments—an idea reminiscent of Plato's classic suggestion that we should have "philosopher kings".[57] On the other hand, if the best way to get truth-tracking moral beliefs is through direct perception or emotional reaction, we might instead give the extra weight to uneducated, "earthy" people whose genuine intuitions are the most untainted by confusing doctrines and obscure theories. Unfortunately these sets are complements of one another, so there is little more I can say about them without taking sides in the meta-ethical question, which is outside the scope of this dissertation.

However, both sides of the meta-ethical dispute would agree about some things. For example, we should give extra weight to the views of people who have detailed experience of the problem at hand and of the policies which are being considered as solutions, as opposed to people who have only a vague idea of what the decision is about. This is because people with a clear image of *what* they are evaluating are likely to evaluate it better, whether their evaluation involves theoretical reasoning or emotional reaction. In more concrete terms: if Marie Antoinette pictures a grain shortage as something which would affect the availability of bread but which would leave cake and other baked goods unaffected, Marie Antoinette is probably not the person who should be judging the importance of averting a grain shortage—regardless of what the correct way to go about such a judgment may be. Note, incidentally, the connection between this idea and Section 2.3.1's claim that we ought to expose people to broad experiences for the sake of improving the accuracy of their moral judgments. From the assumption that broad experience promotes accuracy, it follows that the Neutral Policy will be to encourage people to gain broad experience *and* to listen more closely to people with broad experience than to people without it.

In addition to giving extra weight to the preferences of people who are more likely to have *accurate* moral beliefs, we should also give extra weight to the preferences of people who are more likely to be *motivated by* their moral beliefs. If one person appears to be sincerely pursuing his conception of the moral good, while another person is acting hypocritically, we should care more about helping the former than the latter. I fear that I have no special advice about how to recognize sincerity or hypocrisy; that is an empirical matter.

In cases where a tradeoff is not only between advancing one person's goals over another but between giving one person the power to pursue his goals or giving another person such power—e.g. the kind of tradeoff involved when deciding how to distribute primary goods—we also need to look at how likely the people in question are to use the aid efficiently. Some people are more prone to perform counterproductive actions—counterproductive with respect to their own goals, I mean—than others are. Here we can consider general intelligence as well as technical expertise. Our focus should be on getting resources to the people who are least likely to waste those resources, i.e. the best and the brightest.

A factor worth special mention in the context of both informedness and expertise is a person's position in time. I discussed this back in Section 3.1, distinguishing "give more weight to future people than to past ones because we care more about the former" from "give extra weight to future people than to past ones because the former are more likely to have truth-tracking views", but the latter issue is worth revisiting. At least since the invention of technologies such as the printing press and the digital computer, it has become easier for our civilization to gain knowledge than to lose it: our ability to make millions of copies of our most valued information ensures that future generations will have access to that information—as well as access to whatever new information is discovered in the future. The result is intellectual progress. Over time, people grow better informed about moral issues, descriptive questions, and technical questions about how to achieve various goals. I could be mistaken—perhaps civilization is more cyclical than progressive, and I just happen to be observing a forward-moving part of the cycle—but if I am correct, giving extra weight to people who are more likely to have moral success will entail giving more weight to people the later their position in the history of civilization. This is important in part because "position in history" is a relatively easy property to measure in comparison to things like time "time spent considering a given issue", and policies which discriminate on the basis of position in history may be relatively easy and uncontroversial to implement in comparison to policy which discriminate on the basis of things like education level.

This entire elitist discussion of giving extra weight to the goals of some people over the goals of other people may strike some readers as surprising, at odds as it is with most major moral theories' commitment to human equality. So this might be a good time to remind the reader of what the Neutral Policy *is*. If it were a theory of objective goodness which ascribed intrinsic moral goodness to human well-being, grounded perhaps on universalization-based reasoning such as "I want others to take my well-being into account, so I shall take others' well-being into account", it would have to explain how accidental properties, such as how intelligent or well-informed a given person happens to be, could possibly influence the intrinsic moral significance of that person's well-being. But it is not such a theory. Rather it is a policy based on the *instrumental* goodness of goal-fulfillment, grounded on the idea that people sometimes base their goals on accurate moral considerations. It cares about goal fulfillment because of the impact goal fulfillment has on the rest of the world. So it should be no surprise that it treats some people as different from others, since we know perfectly well that some people play larger roles in history than others do.

It is tempting to go one step beyond evaluation of the reliability of *people* and look at the views themselves. Obviously, we are no longer engaged in a theory-neutral argument if we say "such-and-such view's popularity should be discounted on the grounds that it is less plausible than the views with which it competes"—that would be a theory-based judgment of which views to support. However, possibly we *can* say "such-and-such view's popularity should be discounted on the grounds that it is the sort of view people would tend to embrace for bad reasons". After all, some views will naturally be more popular than others, regardless of their respective truth values. Simple views tend to be easier to hold in one's mind, and easier to communicate to others, than complex views—giving them a competitive advantage. In other words, we should expect simple views to be relatively popular, even when false. A similar competitive advantage is enjoyed by views which elicit strong emotional reactions such as fear or excitement, since a person is more likely to pay attention to such a view than to a less sensational one. A subcategory of this case is wishful thinking: people may be particularly inclined to give undue weight to beliefs which elicit happiness or hope. Still other views confer pragmatic advantages on their believers. For example, if a person asserts "the occupation for which I am trained is especially important to society", he will be able to practice that occupation more enthusiastically; he will also have a ready-made rationalization for non-moral, self-interested behavior that generates conveniences for him at others' expense; and, if he can convince others to share the view, he may even get their acceptance and support for such behavior. Or if he asserts "such-and-such organization"—it might be a political party, a particular branch of government, a non-profit organization, a religious organization, or what-have-you—"is doing morally valuable work", he will find it easier to make friends with members of the organization in question. And so on; there are many views which people have an incentive to believe, to pretend to believe, or to try to get others to believe.

If we could, with a reasonable degree of objectivity, identify beliefs with non-truth-related competitive advantages—and I take it that, at present, it *is* at least slightly easier to reach a consensus on the psychological question "which of these moral views is a person most likely to endorse, assuming that he has no epistemic reason to favor one over any other?" than on the moral question "which of these moral views, if any, do we have most epistemic reason to favor?"—then we would be justified in discounting them. After all, my above claim that a popular belief is, all else equal, more likely to be true than an unpopular one was based on the claim that a true belief is, all else equal, more likely to be popular than it would have been if it were false. So if we know that one view was quite likely to be popular even if it is false, observing that it is indeed popular—assuming that it is not fantastically *more* popular than expected—gives us less reason to conclude that it is true than we would have if it were instead a view which was highly unlikely to be popular if false.

It may seem as though I am committing the genetic fallacy here, both in my discussion of giving extra weight to some people's goals and of giving extra weight to some goals based on their content: the genetic fallacy is inferring that an idea is false, on the basis not of evidence against it but rather of where it came from. True ideas can come from non-truth-tracking sources, as the proverb "even a stopped clock is right twice a day" reminds us. However, *dismissing* ideas based on their origins is not what I am doing here. I am not saying that we should believe people's goals to have negative moral value, and actively seek to frustrate them, when we suspect that those people formed the goals in non-value-tracking ways. All I am saying is that we should ascribe those goals *less* positive moral value than we do goals which are more probably value-tracking. A consequence which a hundred stupid-seeming people are pursuing for what appear to be non-moral reasons, or on the basis of what appear to be seriously biased moral judgments, is nevertheless more likely—all else equal—to be a morally good consequence than one which nobody at all is pursuing: much the way, if we see that a clock which is *probably* stopped but *possibly* in good working order reads "2:00", and we have no other clue about what time it is, we should conclude that it is *slightly* more likely that the time is 2:00 than that it is some arbitrarily-chosen time such as 3:15. However, the goal favored by a hundred stupid-seeming people for apparently bad reasons, while *slightly* more likely to be good than a completely arbitrary goal, is less likely, all else equal, to be good than a goal which a hundred bright-seeming people are pursuing out of what appear to be sound reasons: given a dispute between a clock which is probably stopped but possibly in good working order, and a clock which is possibly stopped but probably in good working order, one should trust the latter. This is not *lowering* one's credence in views as a result of suspecting that they came from non-truth-tracking processes, only *failing to raise* one's credence in them the way one would raise it if they were the product of more reliable processes. So it is not an instance of the genetic fallacy.

The take-home message of this section is this: in conflicts between fulfilling the goals of one group of people or fulfilling the goals of some other group of people, the Neutral Policy will be to side with the larger group, all else equal. In that sense, it is democratic. However, the "all else equal" clause is playing a very large role here. All else equal, the Neutral Policy will also, like utilitarianism, side with whichever group has the most interests at stake. Unlike utilitarianism, it will also side, all else equal, with whichever group is most independent-minded, best informed, most sincere, or most competent, or with whichever group has goals which look least likely to be the products of biased thinking. Numbers are just one of many considerations which it would have us take into account when resolving conflicting claims on our beneficence.

3.4 – Liberty

I will now turn more directly to the kind of political system which would be favored by the Neutral Policy. We saw in the previous section that the Neutral Policy, when faced with a conflict between the morally-motivated goals of two groups of people, will be to favor the larger group—all else equal. That might give the impression of saying "we have a theory-neutral reason to establish, as our political system, absolute democracy: allowing any and all issues to be decided by majority rule". However, I think this would be a mistake. I will argue in this section that the political system we have most theory-neutral reason to establish is actually a *liberal* democracy: one in which individuals and minority groups have rights—to liberty, property, and political participation—which cannot be overridden by a majority vote. The Neutral Policy should be to respect and protect those rights, even when this means going against the will of the majority.

My focus here will be on distributional questions: what should be done with society's resources? These include natural resources like petroleum, manmade resources like trucks and factories, and human resources such as people's time and effort. We can imagine groups of people with shared goals forming and lobbying to be able to dedicate all these resources to their favorite cause: for example, a group of progressives who want all of humanity's wealth and energy spent on the maximization of human flourishing, or a group of conservatives who want it devoted to safeguarding human life and traditional human lifestyles. Perhaps another group would consider all of these goals important and is pursuing a balance between them. Absolute democracy would assign all the resources—even power over how others use their labor—to whichever groups could form a majority coalition. The form of liberalism which I will advocate, on the other hand, allows every group to control some material resources, and to do what it wants with its own members' labor. I think we have a theory-neutral reason to establish the latter sort of system.

One note before I begin: when I write of the allocation of resources to groups or individuals, keep in mind that this a question about who *controls* their use, not about who will ultimately *consume* them. For example, if we decided to allocate all food resources to a single philosopher king, that would not mean that everyone else would *starve* due to having no food. It would instead mean that the king would decide how the food would be distributed—e.g. equally, based on need, based on efficiency, based on their willingness to obey him in other matters, or whatever. If we decided to allocate *all* resources to the philosopher king, then he would also be controlling the distribution of other materials, as well as dictating what everyone would do with their time.

3.4.1 – Against Concentration of Power

This section will argue against concentration of power. I do not consider it desirable for any individual, nor any group of individuals with shared goals, nor even a coalition of such groups, to have control over all of society's resources. I shall argue that even the best-chosen government should not have the power to take individual's property or control how individuals use their labor.

First we have to ask how well-chosen a government could be. We saw in the previous section the many *departures* which the Neutral Policy would make from pure one-man-one-vote democracy. It would, at least ideally, give extra weight to groups composed of independent-minded, well-informed, morally-sincere, highly-competent, and non-biased members. Implementing this ideal, however, would be tricky. What would we do, have some sort of qualification exam at the voting booths which would determine the number of votes each citizen was permitted to cast? Aside from the ridiculous expense that would entail, and the unpleasant memories evoked of racist Jim Crow voting restrictions, there is also the problem of who would design the exam. People who sincerely believe that their values are the morally true ones, and that they have a moral obligation to fulfill those values as best they can, will naturally be tempted to skew the exam to favor people who think like themselves. Any actual voting system, then, is going to fall far short of the description from the previous section. The winning group or coalition may be *slightly* more likely to be pursuing the right goals than the nearest runners-up are, but *only* slightly.

Suppose we had just two options for how to distribute society's resources: allow all of them to be allocated to serving the most popular set of moral goals, or else divide them evenly between the *two* most popular such sets. "Popular" here, let us suppose, is measured by some sort of voting system and is a *very* rough indicator of which goals are value-tracking; and let us also suppose that the runner-up is not far behind the leader in terms of popularity. So let us estimate that both groups are approximately equally likely to be pursuing the correct goals. The question, then, is whether it is better for all resources to be under the control of one group or for each of the roughly-equally-value-tracking groups to control half of the resources. I will argue that the latter is better, that concentration of resources toward a single purpose is to be avoided when there is no consensus around that purpose.

Concentrating all resources into one group's hands gives us a chance that all of them will end up dedicated to morally correct priorities, but also a chance that all of them will be spent on goals which are not morally significant after all. Dividing the resources among two groups more-or-less eliminates the best-case scenario—the new best case is that only half of them will be spent on the correct goals. On the other hand, it roughly doubles the likelihood that the worst-case scenario will be averted and that *some* resources will be spent on the correct goals. So this is in part a question about moral risk-aversion: in the case of moral uncertainty, it more important to reduce the likelihood of a morally terrible scenario or to strive for the morally best scenario?

I fear that I do not have a good answer to this question. There is something to be said for a maximally-risk-averse maximin approach which attempts to steer clear of moral disasters even if this guarantees a morally mediocre outcome. There is also something to be said for a "maximax" approach which tries to give us at least some chance of behaving exactly optimally. Personally I think the most plausible approach is the risk-neutral expected rightness strategy described in Section 1.1.1, which, if used in sufficiently many independent contexts will produce better results than either of the other two—"sufficiently many" being however many it takes for the law of large numbers to allow the steady trickle of mediocre results produced by a maximin strategy to outweigh whatever rare disasters that strategy averts, and to obliterate the maximax strategy's chances of perfect success. But I do not really want to argue for moral risk-neutrality here; it is too far outside the scope of the aims of this dissertation. Instead I will simply say: I am *assuming* moral risk-neutrality, without arguing for it. I am assuming it for simplicity, and because it is in the middle of a continuum of reasonable positions stretching from extreme risk-aversion, through moderate risk-aversion, through moderate risk-seeking, to extreme risk-seeking. If the reader does not share my stance of moral risk-neutrality, he or she should adjust my view to accommodate *his* stance—if he or she is morally risk-averse, the adjustment will entail viewing the case against concentrating resources into a single group's hands as even stronger than I am representing it, whereas if he or she is morally risk-seeking, it will entail viewing the case as weaker than I am representing it. At least my discussion will give him or her a jumping-off point from which to make those adjustments—one which requires *less* adjustment than if I had assumed a level of moral risk-neutrality on the opposite side of the continuum from his or her own, rather than assuming a level in the middle of the continuum.

Given my assumption of moral risk-neutrality, my version of the Neutral Policy will be indifferent between an X% chance of fulfilling true moral values to degree Y, and a 2X% chance of fulfilling true moral values to degree Y/2. However, that is not the choice which is at hand. The choice at hand concerns not degrees of fulfillment but of resources: would we rather have an X% chance of dedicating quantity Z of resources to the pursuit of true moral values, or a 2X% chance of dedicating Z/2 of resources to the pursuit of true moral values? The distinction matters: the degree to which a value is fulfilled is *not* linearly proportional to how many resources are dedicated to pursuing that value.

For example, consider the law of diminishing marginal returns, familiar in economics—or what is known in folk proverb as the "80-20 rule", that eighty percent of an effect can be achieved with only twenty percent of the work. It applies just as much to production of moral goods such as well-being or equality as it does to the production of widgets and gizmos. Suppose that followers of utilitarianism had a billion dollars' worth of goods and labor to spend promoting utility. They could spend it where it would be most effective—say, helping relieve the suffering of people with easily-treatable diseases who could be cured for mere pennies, or making investments in education and infrastructure aimed at ending poverty entirely in the long term. But now suppose they had a second billion dollars. The most easily-treatable sufferers would already have been treated, and the most promising infrastructure investments would already have been made; so the second billion would have to be devoted to harder-to-treat causes of suffering or less-promising investments. It seems a safe guess that the expenditure second billion will result in a *smaller* increase in utility than the first billion did. So if utilitarianism turns out to be the true moral value system, a 2X% chance that one billion dollars will be spent in its pursuit is preferable to an X% chance that two billion dollars will be spent in its pursuit.

However, I do not want to rest much weight on this guess. The law of diminishing marginal returns will not always the dominant factor. There might be economies of scale: shipping a *full* container of food or medicine to a given needy region is likely to be more cost-effective than shipping a *half-full* container of food or medicine to that region, but is only an available option if the donors have enough money to fill the entire container. Also, sometimes it is necessary to spend resources to *create* opportunities before one can spend them *exploiting* those opportunities: "teach superior agricultural methods" might be an important and cost-efficient project, but it has "develop superior agricultural methods" as a prerequisite, and being able to afford the former is useless unless one can also afford the latter. These are just two of many possible "tipping points" where the marginal effectiveness of resources *increases* as the total available resources pass a given threshold. On average, the law of diminishing marginal returns should still hold—after all, each tipping point is two-sided: the marginal utility of additional resources is relatively high when one has a not-quite-full container, but is relatively low when one has a single full container and nowhere near enough goods to fill a second container—but there is no guarantee that it will hold at any particular level of resource availability. So it may turn out that two billion dollars *can* produce more than twice as much utility as one billion dollars can; I do not think it *probable*, but it is possible.

Furthermore, it very much matters how exactly we map utility to value. Consider the following two theories of moral value:

> *Moderately-risk-averse total utilitarianism* holds that increasing the world's total utility by a factor of three increases its total moral value by a factor of only two.[58]
>
> *Moderately-risk-seeking total utilitarianism* holds that increasing the world's total utility by a factor of only two increases its total moral value by a factor of three.

Both of these theories agree that more utility is always better than less utility. However, the former holds that the *marginal* value of utility decreases with quantity, judging that bringing the first X utils in the world is as important as bringing about the next *2X* utils is; the latter holds that the marginal value of utility *increases* with quantity, judging that the step from 0 utils to X utils is only half as important as the step from X utils to 2X utils. As a result—and assuming a "maximize expected rightness" paradigm—the former would discourage us from making fair bets with dollars even if the marginal utility of dollars were constant; for example, in a choice between an option in which there is a

100% chance of one million dollars being spent on utility-promotion and an option in which there is a 50% chance of two million dollars being spent on utility-promotion but a 50% chance of nothing being spent on utility-promotion, it would tell us to choose the safe option. The latter, on the other hand, might well tell us to choose the risky option in that situation. It might say "choose the risky option" even given diminishing marginal utility of dollars, so long as the marginal *utility* of dollars diminishes slower than the marginal *value* of utility increases.

Given that such theories can be formulated, the conversion from dollars to moral-goodness-according-to-utilitarianism is even shakier than the conversion from dollars to utility was. The law of diminishing marginal returns may still hold *on average* across a majority of possible moral theories—if we can make sense of talking about "majorities" when there are an infinite number of possible variations along this dimension—but it cannot be expected to hold of *every single theory*. Still, this point is reinforced: a theory's degree of fulfillment does not necessarily increase linearly as the resources controlled by its followers increase. That will be enough for my present purposes. Call a value system "risk-averse with respect to resources" to the extent that having half the world's resources dedicated to its fulfillment gives it a higher expected level of fulfillment than a 50% chance of having all of the world's resources dedicated to its fulfillment. Call it "risk-seeking with respect to resources" to the extent that it has a higher expected level of fulfillment in the latter case.

Let us now return to the question of whether is it better to dedicate available resources to the fulfillment of one highly-plausible value system, or to divide it up and dedicate half of the resources to one highly-plausible value system and half to a second highly-plausible value system. Setting policy in advance, we do not know whether these value systems will be, on average, more risk-averse or risk-seeking with respect to resources. If they are risk-averse with respect to resources, dividing available resources among followers of both will result in a higher expected overall degree of fulfillment than giving control of all available resources to followers of just one. How about if they are risk-seeking with respect to resources? One might think that expected overall fulfillment would be maximized in this case by concentrating the resources into a single group's hands, but this is not necessarily so: if it is divided equally, and if the two groups of followers are free to do what they want with their share of resources, they can always make a fair bet with one another—e.g. betting it all on a coin toss—and end up with just as much expected success as if the system had arbitrarily picked one. So if two value systems are risk-averse with respect to resources, dividing control of resources across both of them is better than concentrating control in just one of them, whereas if they are risk-seeking then both options are equally good. So if we do not know whether the value systems will be risk-seeking or risk-averse—let alone if we think the law of diminishing marginal returns makes the average value system at least slightly more likely to be risk-averse than to be risk-seeking—we should choose "divide the available resources among both groups, allowing them to agree to gamble with it if they prefer that option" rather than "concentrate all available resources into the hands of a single group".

A further, related advantage of division involves not mutually-beneficial gambles but mutually-beneficial trades. I remarked earlier that sometimes different value systems have different amounts at stake in a given situation. Perhaps a particular good is much more useful for fulfilling one than for fulfilling the other. Perhaps a particular individual's action or a particular piece of public legislation would have a much bigger effect on one value system's level of fulfillment than on the other's. Concentrating all wealth and power into one group's hands means that it will get even those bits of wealth and power which would be much more useful to the other group. On the other hand, if we divide resources among both groups and allow trades, the two groups can make trades until each controls the bits which it regards as most useful. Alternatively, we could try to take into account what each group cares about when we make the initial division—the "trade" element is one easy way to do this but is not essential. What is essential is that if the two value systems do not have the same metrics for quantifying resources, so our options are not "give one group all available resources, or give each of two groups half the available resources", but rather "give one group all available resources, or give each of two groups an amount of resources which *it regards as* more than half of what is available".

Of course, not all possible trades will take place. If resources are allocated to followers of both of the two most popular moral theories, some of those resources will end up being used at cross purposes. For a familiar example, consider two advocacy organizations on opposite sides of a given issue such as abortion or gun control, each spending money on advertisements to try to influence public opinion or on lobbying to try to influence members of government. To the extent that the efforts cancel one another out, this expenditure has been wasted. It would be better if the two sides agreed to a truce and instead spent their resources on goals they agree about—e.g. preventing unwanted pregnancy, preventing theft of guns from their rightful owners, etc. But would it be better if one side had all the resources, so its lobbying efforts to further its agenda could go unopposed? I think not. With respect to the average level of fulfillment of the two theories, any resources spent in a way which furthers one side's agenda while setting back the other sides' agenda to the same extent is wasteful: it does as much expected harm as expected good. The wastefulness of such zero-sum expenditures may become more *visible* when both sides are making such expenditures and cancelling each other out, but such expenditures *are* wasteful even if only one side has the means to make them. If anything, dividing resources at least gives *some* bargaining power to both sides, creating a *chance* that they will recognize the waste of working at cross purposes, compromise, and start working toward shared goals rather than zero-sum ones.

So far I have argued that, all else equal, our Neutral Policy should favor dividing resources among pursuers of the two most plausible moral theories, rather than concentrating those resources in the hands of pursuers of the single most plausible moral theory—provided that the two theories are, as far as we can easily measure, roughly equally plausible. A similar argument, *mutatis mutandis*, would show that dividing it among three groups with distinct plausible value systems is better than dividing it among only two groups, and so on. At the extreme, if we had no faith at all in the truth-tracking-ness of opinion polls—for example if we thought that almost everyone formed moral value judgments based on blind trust in someone else's judgments, rather than forming them independently, so thought that the overwhelmingly most likely explanation for why a billion people accepted a given theory is that one person came up with it, perhaps by recognizing it as true or perhaps by making a mistake, and the rest blindly copied his view—we could divide resources equally among *all* value theories with at least one subscriber.

In reality, assuming that we do think that a view with a billion followers has *something* more going for it than a view with only a hundred followers—even recognizing the possibility that the latter's followers are more independent-minded, better informed, less biased, or whatever, in some way that we have failed to detect—we will want to seek a compromise between absolute concentration of resources, into the hands of followers of the single most popular value system, and extreme division of resources, to be spent toward the fulfillment of every value system under the sun. I shall now discuss one such possible compromise.

3.4.2 – For Liberalism

The conclusion of Section 3.4.1 was that it is subjectively better to allocate control over *some* resources to followers of many plausible value systems, rather than giving full control to followers of a single value system. This leaves open, however, the question of which value systems count as sufficiently plausible, and the details of the distribution. What I want now is to discuss one particularly salient possible policy of distribution, which I shall call "liberalism", due to a *partial* resemblance to classical liberalism. However, please note that—unlike classical liberals—I am concerned with it *qua* allocation of resources, not *qua* objective moral theory.

> *Liberalism* is the policy of allocating to each individual the freedom to dedicate his own property to the pursuit of his own values, where such property includes, first, his own body and his own labor insofar as he has not voluntarily given or traded them to others, second, anything voluntarily given or traded to him by its previous owner, and third, anything produced by altering or combining bits of his existing property.

I leave unspecified the answers to questions about how unowned goods—natural resources, goods abandoned by their former owners, etc.—should pass into ownership, as well as questions involving the restitution of wrongfully appropriated property. I also leave unspecified the answers to questions about where one person's freedom ends and another's begins—e.g. in cases where one person doing what she wants with her property interferes with another person doing what *he* wants with *his* property. There are multiple ways to resolve those questions, any of which would be consistent with the general idea of liberalism as I am envisioning it. I also leave unspecified, for now, the answers to paternalistic questions about when society may intervene in someone's behavior if that behavior is foolish or akratic and so *not* advancing that person's own values.

As a distribution of resources to followers of different value systems, liberalism is a compromise between allocating all of it to be used in pursuit of the single most popular value system, and allocating equal amounts to every plausible value system. On the one hand, each value system with at least one follower gets allocated *some* resources—namely the bodies, labor, and personal possessions of its own followers. So many value systems will get their very highest priorities fulfilled, as discussed in Section 3.4.1. Since liberalism as defined above permits people to make trades and to keep the products of those trades, it also reaps the advantages discussed in Section 3.4.1 of permitting mutually-beneficial exchanges. On the other hand, relatively popular value systems will tend to get *more* resources than relatively unpopular value systems will, by virtue of having more followers—at least provided that individuals have roughly-equal property, or at least that inequalities in property do not systematically correlate to which moral theory different people believe in. So if, as suggested in Section 3.3, popularity turns out to be a measure of value-tracking-ness, the distribution will have responded to popularity at least somewhat.

Of course, many possible allocations of resources could display such virtues. So the task for this section is to argue that there are specific advantages of liberalism that are not shared by other possible allocations. This section will *not*, of course, be arguing that individual liberty or property rights are intrinsically good, or are "natural rights", or anything of that sort; such arguments would be theory-based and so outside the scope of this dissertation. Instead, my claim is simply that the Neutral Policy—the set of rules which we have a theory-neutral reason to establish and so *should* establish in the absence of countervailing theory-based reasons—includes respect for individual liberty and property.

The first specific advantage of liberalism is its ease of implementation. Many possible allocations of resources—e.g. ones of the form "allocate an equal share of the world's resources to each value system with at least N followers"—would require careful polling to find out how many people believed in each value system. The situation would be complicated by the incentives people would have to vote strategically rather than sincerely, e.g. supporting a lesser evil because their true values were below the threshold for representation. So polling, and especially accurate polling, would be expensive; it would consume resources that could instead have been used for advancing plausible values. In contrast, liberalism as defined above does not require any sort of polling to implement; nobody except the individual himself needs to know what values he supports.

A second advantage of liberalism has to do with enforcement. Allocating someone's labor—or whatever he can produce with that labor, in private—to the advancement of values with which he disagrees is difficult. He will try to find ways to cheat and advance his *own* values instead, or to slack off and advance non-moral goals. Preventing this will require supervising his work and having some sort of enforcement mechanism in place, both of which will consume resources. And even if he can be coerced into *attempting* to cooperate, it is unlikely that his work will be very efficient. It will be hard for him to muster enthusiasm for his assigned task, and for many tasks enthusiasm is useful. Furthermore, he may not fully *understand* the goals he has been ordered to pursue, which may lead him to make mistakes. These problems can be reduced, albeit not eliminated, by liberalism: although still tempted to pursue non-moral goals, people can be expected to feel *more* motivation to pursue their own values than to pursue someone else's values; and although many people will undoubtedly hire their labor out to others, they will have at least done so voluntarily and will have a wage incentive to continue being useful to their employers.

I should not exaggerate these two advantages, since they are mitigated somewhat by the questions I left unspecified in my formulation of liberalism. We will probably want to use a democratic process to deal with the problems of distributing unowned resources and of resolving conflicts.[59] Some degree of paternalism—intervening when people's behavior becomes too unwise, too inconsistent with promoting even their *own* values—is probably desirable, and will require enforcement. Maintaining the rules, preventing people from enslaving one another or appropriating one another's property, will also carry an enforcement cost. Incidentally, I assume that one of those rules needing to be enforced will be some sort of taxation system used to fund all these other activities. Taxation need not be construed as a violation of the liberal ideal, but rather as a fee people pay society in exchange for benefits such as dispute resolution and law enforcement. Notwithstanding all these expenses, I do think polling and enforcement costs can be expected to be *less* in liberal sociopolitical systems than in ones significantly diverging from liberalism.

I now turn to a third advantage of liberalism, the one which I think is the most important from the perspective of theory-neutrality and which depends on features of theory-neutral reasons beyond mere cost-avoidance. It is this: liberalism promotes moral and scientific progress, helping people to have the information they need to choose *better* projects. Many of my arguments will be familiar from classical liberal political philosophers—especially John Stuart Mill[60]—but it will be worthwhile to reiterate them here in order to show that they are not theory-based: even if we reject Mill's utilitarianism, his argument for liberty still stands so long as we have credence in any at-least-partly consequentialist moral theories.

I say that liberalism promotes "moral progress". What I mean is that it increases the availability of information about which things are intrinsically morally valuable, and about the extent to which those things are intrinsically valuable. So the Theory-Neutral Reason to Promote Recognition will, all else equal, advocate that we select a liberal policy. Or, to put it in the framework of the discussion from Section 3.2: selecting a liberal policy will cause relatively-value-tracking goals to be adopted, which will in turn increase the fulfillment of those goals; so selecting a liberal policy will increase the overall fulfillment of potential value-tracking goals.

How does liberalism promote moral progress? It does so in several ways. One is by encouraging the production and dissemination of moral arguments. By tying the amount of resources which will be directed toward fulfillment of a given value system to the number of subscribers to that value system, it creates a moral incentive for people to try to spread their values: if one believes a given consequence to be good, one will want other people to spend their labor and other resources in pursuit of that consequence. But by assuring people control over their own bodies and property, it prevents coercive efforts to recruit others to one's cause, which leaves the option of *persuading* others to adopt one's values. So it creates an incentive for people to invent and publish new *arguments* for their views. And by giving people control over their own labor, at least while pursuing their moral values, it gives them the freedom to act on that incentive. Even if a new value system has only a single subscriber, he will be able to explain *why* he subscribes to it, and try to win new followers. Something like this may not just be useful for moral progress but in fact be essential, since it is almost inevitable that a new discovery will start out believed only by its discoverer, and only later spread through the population—*if* he is free to spread it and has a reason to do so.

There are several comments to be made here. First is a reminder of what is meant by "value system". We are not concerned with a person's self-interested values, nor with his agent-relative or non-consequentialist moral beliefs. Instead we are concerned with his agent-neutral, consequentialist moral beliefs—ones with the structure of "it is right not only to fulfill these values yourself, but to cause someone else to fulfill them as well". We are concerned with them because our goal is to act rightly *ourselves*, not necessarily to cause other people to act rightly. Also, they are the type of value which people will have a moral incentive to spread in a liberal system. There is not necessarily any incentive in the liberal system for citizens to create arguments for the agent-relative duties in which they believe, but this is not a problem for the Neutral Policy since we—policy-makers—have no theory-neutral reason to be concerned with helping people obey their agent-relative duties. If a further refresher on this is needed, see the discussion from the start of Section 2.3.

A second note: I should explain *why* I think production and dissemination of moral arguments facilitates moral progress. The incentive here, after all, is for producing *persuasive* arguments, not for producing *sound* arguments. The production of unsound but persuasive arguments does not facilitate moral progress, since beliefs formed on the basis of unsound arguments are non-truth-tracking. However, I think soundness is *one* factor which can make an argument persuasive. Other factors—e.g. brilliant rhetoric—can in principle be added to a sound argument as well as to an unsound one. To use an example from commerce: I take it that automobile dealers have discovered that showing advertisements in which beautiful women are posed beside their product helps them sell more of that product. This might seem like an unfortunate distraction from issues like how well the cars actually run. However, a pretty woman can be posed beside a good car as easily as beside a lemon, so *all* the auto dealers use this technique. So the dealer selling higher quality cars still has an advantage in the marketplace: not only are his cars, like everybody else's, associated with beautiful women, but they also run well. Similarly, we can hope that the most persuasive moral arguments, *in addition* to employing various tools of rhetoric, will also be sound; and that therefore incentives to produce arguments which are as persuasive as possible will also lead to the production of sound arguments. Recall my discussion from Section 2.2.1, in which I said that we should have at least some credence that moral argumentation tends to be truth-tracking—especially given the costless assumption that we have *some* way to separate moral truths from moral falsehoods. If that discussion is right, then we should be in favor of incentivizing the production and dissemination of moral arguments.

Furthermore, it is not necessarily true that unsound arguments are useless. John Stuart Mill argues that even if an idea is wrong, we can gain intellectual benefits from considering it and learning *why* it is wrong. Attempted counterarguments to a widely accepted view, even if they ultimately fail, can pioneer new methods of reasoning which later bear fruit—in addition to being good training for our general reasoning abilities.[61] Likewise, our responses to attempted counterarguments can result in a deeper and more nuanced understanding of the view we are trying to defend.[62] Also, just knowing that a view has been challenged and survived the challenge is a kind of argument *for* that view—since a view is more likely to be true if it has survived many attempts at falsification than if it has never been tested—and so in a way a collection of unsound arguments for incorrect value systems can constitute useful evidence for the correct value system.[63]

Production of moral arguments is not the only way in which liberalism can facilitate moral progress. After all, freedom does not just mean freedom of speech, freedom to argue for one's values; it also entails freedom of action, freedom to pursue one's values. For people to exercise that freedom may help everyone determine whether those values are truly worth pursuing. Section 2.2.1, as an alternative to its account of how we might recognize moral truth by way of argument, suggested that we might also be able to recognize moral truth simply by examining moral reactions to various situations. Such reactions can, the thought goes, identify moral values which had previously been overlooked; some goods may be undreamt-of until one sees their presence, or taken for granted until one experiences their absence. Certainly this happens with subjective evaluation of outcomes; we think we want one thing, like career advancement, and only later, when we have achieved that goal but still feel unhappy, do we realize that something else, such as friendship, was really more important. It seems plausible that an analogous process could happen with objective values as well. Something which seems like a morally good idea at the time—for example, the eugenics craze of the early twentieth century, many of whose proponents truly believed in what they were doing, comes to mind—can produce feelings of guilt and revulsion in hindsight. Experiencing the effects of a mistaken ideology is an expensive way to learn a moral lesson, but sometimes it may be the only way. Liberalism, by allowing individuals, and voluntary associations thereof, the freedom to pursue their own values, gives us the opportunity to see what happens when those candidate values are implemented on a small scale. Our reactions to that implementation may well constitute truth-tracking information.

This point too, of course, is straight out of Mill. He writes, "As it is useful that while mankind are imperfect there should be different opinions, so is it that there should be different experiments of living".[64] It is worthwhile to unpack the details of this simile. One of the advantages of free speech is that good ideas—sound arguments—can sometimes be recognized as such, and so will tend to spread if expressed at all; but if there is a perceptual component to the way in which we get moral information, then good actions can also be recognized as such. Likewise, just as unsuccessful arguments were still useful in some ways, unsuccessful "experiments of living" could be too—both by teaching us specific lessons about how the particular experimental outcomes in question turned out to less good than expected, and also by teaching us more general lessons about the best way to conduct such experiments and to avoid overconfidence about the quality of their outcomes.

I claimed that liberalism contributes not just to moral progress but to scientific progress. What I have in mind here is the production of two types of scientific information: descriptive and technical. By "descriptive information" I mean, simply, information about what is true in a given situation; the kind of information useful in helping to determine whether a given outcome counts as fulfilling a given morally-motivated goal. For example, suppose that one is opposed to there being pain in the world—perhaps due to utilitarian moral convictions, perhaps due to instinctive sympathy, or perhaps just as an aesthetic preference. Before one can do much about it, one needs to know which situations include pain and which do not. Do other people have minds and experience pain, or are they just zombies or figments of the imagination? How about lower life forms—are we guilty of anthropomorphism if we impute pain to a fish which is flopping around after being removed from water, or is it actually suffering? The best way to answer such questions is probably by collecting data—observational and experimental—on nervous systems, behavior, evolutionary history, and so on. Once we have the data, we can then consider various models and metaphors to help us understand what is happening, and perform an inference to the best explanation. Liberalism can help with both steps; replicable data and useful models can both be spread through the marketplace of ideas in the same way that sound moral arguments can—perhaps even better, since we have a clearer idea of how to evaluate scientific claims than of how to evaluate moral ones.

By "technical information" I mean information about how to bring about particular results. It would be no use to have the right morally-motivated goals, regardless of whether we could recognize their fulfillment, if we had no idea which of our actions would lead to their fulfillment. We would be no more likely to further our aims than to do something counterproductive. Descriptive models are helpful here, so this category of information is not entirely distinct from the preceding one; liberalism improves our access to technical information by improving our descriptive models. However, technical information can also arise in the absence of descriptive models, through the simple process of trial and error—people perform an action, see whether the results are satisfactory, and then either repeat it or perform a different action next time. Just as the increased freedom of action in liberalism permits moral experimentation, allowing us to experience and evaluate various consequences, it also permits technical experimentation, allowing us to observe *which* consequences arise from which types of actions. There is also room for serendipity—an action whose effect is not the one its own agent wanted can be noticed by another agent as useful for *his* purposes.

Incidentally, we can now see why Lloyd Thomas's defense of liberalism, mentioned in Section 2.3.1, is incomplete. Lloyd Thomas is interested solely in the *moral* information arising from experiments of living.[65] I do not deny that we have a theory-neutral reason to pursue such information, but I do deny that it is the *sole* kind of information we have a theory-neutral reason to pursue. Sound moral arguments, detailed observations, clever models, and experimental results are *all* useful for discovering or implementing correct value judgments, and all result from adoption of liberal policies.

The reader might be wondering whether these things are *exclusively* products of liberalism. Notice that unlike Lloyd Thomas, I do not need to claim that they are. His theory is based on the idea that liberalism is a *necessary precondition* for progress; I only need the weaker claim that liberalism facilitates progress *more than* other feasible social arrangements do. However, even this weaker claim may be non-obvious. Liberalism leads to progress by encouraging argumentation and permitting experimentation. But in principle, a command-based social structure could also permit argumentation and experimentation; rather than issue homogenous commands to everyone, it could command all manner of experiments. *Relying* on it to do so is probably unwise unless its controllers have sufficient awareness of theory-neutral reasons and sufficient humility regarding their present beliefs, but the possibility exists. It could even command *more* experimentation than naturally occurs under liberalism, and thereby foster *more* progress.

Lloyd Thomas calls this idea "command liberalism" and rejects it out of hand: "for exploration of what is of value in itself to have any point, it must be motivated by conviction and inclination".[66] However, that is much too quick. *If* we followed Mill in believing that the only valuable activities were ones valued by their practitioners, *then* it might be true that commanding people to unwillingly perform activities would not significantly help us determine whether those activities were valuable. But such a belief is theory-dependent. If we are not going to commit to utilitarianism, then we should admit that *some* possible intrinsic values could be realized even by unwilling agents, and subsequently recognized as intrinsically valuable.

The possibility of a workable "command liberalism" notwithstanding, however, there are practical problems with commanding people to argue and to experiment. They are identical to the problems discussed earlier with commanding people to produce something those people do not value—for argumentation and experimentation *are*, in a sense, products. The best arguments require both concentration and creativity, and so are difficult to achieve if the arguer is not arguing wholeheartedly.[67] In contrast, freedom of speech allows people to argue for the positions of which they themselves approve, thereby encouraging them to devote their mental resources to making those arguments as powerful as possible. As for experimentation, the best experiments are sensitive to local details—from a remote vantage point, it is easy to miss important factors which will confound or ruin an experiment. It is also easy for an individual to sabotage an experiment involving his own lifestyle, if he does not approve of where he expects it to go; enforcement of commanded lifestyle experimentation would be difficult. Enforcement issues are largely just limitations on information processing, and might someday be overcome by a sufficiently-technologically-advanced totalitarian regime, but for the time being such limitations are reasons for leaving people free to design their *own* lifestyle experiments. Instead of commanding specific experiments, central authorities can do better by offering incentives for innovation—patents, prizes, research grants, and the like—and leaving it to individuals to decide for themselves how to respond to those incentives.

So there are many reasons why the Neutral Policy should involve a generally liberal distribution of limited resources: such a distribution is relatively easy to implement, relatively easy to enforce, and produces morally useful information as a byproduct. I could leave the discussion here, but I have a few more comments to make about the scope of what I have said here, which shall be the subject of the next section.

3.4.3 – Implementation

As I have described it, liberalism is more an ideal than an attainable reality. A socio-political system which protects people's persons and property from appropriation by others does not come for free; it requires a share of the national product to maintain. Furthermore, it will not operate flawlessly: no matter how hard a society tries to protect individual rights, it cannot hope to eliminate all crime and injustice. All we can do is try to promote freedom *as best we can*: implement only those laws and policies whose *net* effect on liberty—that is, their positive effects such as preventing coercion and theft, minus their negative effects such as imposing regulatory or tax burdens on citizens—is greater than all available alternatives. I will discuss here some of the salient features which such policies would have.

When deciding what laws to create, we might appeal to something like John Stuart Mill's "Harm Principle". Mill writes: "the only purpose for which power can rightfully be exercised against any member of a civilized community, against his will, is to prevent harm to others".[68] If we reinterpret "harm" as "inability to use one's labor and property in pursuit of one's morally-motivated goals", the principle follows from the idea that laws should have a positive net effect on liberty: an act, whether by an individual or a government, cannot have a positive net effect on liberty if it has *no* positive effects on liberty at all.

This principle can also be expressed in terms of the language of human rights. We can say that the Neutral Policy includes a right not to have one's lifespan significantly shortened, since premature death would hinder one's ability to pursue his morally-motivated goals. It also includes a right not to have one's autonomy violated via brainwashing, psychoactive drugs, physical brain trauma, etc., or via deliberate deception and manipulation—having one's goals altered from the outside prevents him from pursuing his original goals. Next is a right not to have one's bodily freedom restricted: a right not to be maimed, a right not to be fettered or muzzled, a right not to be enslaved, and a right not to be coerced. Last is a right against not to have one's property taken or damaged without adequate compensation. "Rights" talk is appropriate here because of the absolute structure of these rules: these rights are not to be violated *even when* the would-be violator believes that violating them would have morally good effects overall. That would be a case of the violator appropriating control over resources in the pursuit of *his* values, when those resources should—under the allocation advocated in Section 3.4.2—have been used in pursuit of his victim's values. Instead, the only thing that can override these rights is the protection of someone else's rights.

It is worth taking a moment to dwell on this point. The Neutral Policy, the policy which we have a theory-neutral reason to adopt and to try to get others to adopt, says never to violates others' liberty rights. That is not to say that such "rights" play any role in the true objective moral theory: as always, I remind the reader that I am not taking sides between objective moral theories. It is not even to say that there will not be individual situations in which theory-neutral reasons favor rights violations. Undoubtedly there will. For example, suppose that a legislator comes up with a proposed public works project which would drastically increase the primary goods available to future generations, but which would require the conscription of a million people to construct. Quite possibly the Theory-Neutral Reason to Promote Effectiveness would be in favor of this conscription, and the other theory-neutral reasons would be silent. Nevertheless, I think we should, when setting policy in advance, bind ourselves against taking such actions in the future. In advance, before we know the details of the project and what moral reasons, theory-neutral or otherwise, will support it, we have to be aware of the strong possibility that the person who thinks it morally best to conscript a million others will be mistaken, while the million others who think it morally best that they not be conscripted will not be mistaken. Hence our theory-neutral reason to commit ourselves *now* to following the Neutral Policy *in the future*.

Returning to the listed rights, I want to make a few notes, by way of reminder that the "rights" we have a theory-neutral reason to establish in the law may not be identical to the rights posited by many people's objective moral theories. First, calling the Neutral Policy rights "human" rights is not *precisely* correct: they apply to the beings whose ability to make moral judgments, choose goals based on those judgments, and effectively pursue those goals we have theory-neutral reasons to protect. We saw earlier that these were the beings *capable* of making and acting on moral judgments: this set includes *many* humans, but not necessarily infant or mentally deficient ones; and it might include at least a *few* non-human animals at present, and could someday end up including non-human space aliens, genetically-uplifted animals, or artificial intelligences.

Second, despite their absolute flavor, the rights in question should not really be "inalienable" in any strong sense. If a person consents to be treated in a way which would otherwise be rights-violating—for example, he sells his day's labor to someone else in exchange for wages, or he consents to the administration of a mind-altering drug because he deems that this will serve his goals—there is no obvious reason for the Neutral Policy to intervene and prevent it. It might monitor the situation to make sure the individual in question really did waive his right voluntarily rather than being coerced or tricked into doing so, but if the waiving was voluntary, then the individual must be presumed to be using—in this case, disposing of—his resources in the way that he thinks best. Similarly, if taking away some of the rights of a given criminal is the most effective way to promote rights overall, that too will be permissible under the Neutral Policy. That said, it may be that we should prohibit rights from being alienated for any *long* period of time. If someone wants to sell himself into forty years of indentured servitude, or if the government wants to permanently maim a criminal, we could reasonably object on the grounds that the person the individual *will be* twenty years from now might not pose the same threats to society that he does at present.

Third, the liberty right here is not quite a right of persons to do *whatever they want*. Rather, it is a right to do whatever they think *morally best*. This gives the government some leeway to interfere with individual liberty as commonly conceived. Forcible rehabilitation of drug addicts, conscription of the unemployed into labor battalions, mandatory post-mortem organ donation: all of these could be permissible, as long as the government allows people *who demonstrate a sincere belief that their participation in the program will have morally bad consequences* to opt out. Think "conscientious objector status".

Fourth, there are some conspicuous absences from the brief list I offered above. For example, the Neutral Policy does not include any fundamental right to privacy. There is no obvious reason why the scrutiny of others should stop someone from doing what he thinks morally best. At most, there might be instrumental reasons to protect privacy: for example, to make it harder for the government to enforce bad "victimless crime" laws which allow legislators to impose compliance with *their* values on the citizenry, decreasing citizens' freedom on net; or to make it harder for wrongdoers to accumulate information which can be used for coercive blackmail. Similarly, there is no fundamental Neutral Policy right against cruel or painful treatment—e.g. torture—aside from the general tendency of pain to reduce victims' ability to think coherently, and the potential of using the *threat* of such treatment coercively.

How about "positive" rights? Should the Neutral Policy not only say "do not violate people's lives, autonomy, freedom, or property", but also "actively rescue people who are in danger of losing their lives, autonomy, freedom, or property through no fault of yours"? I think the Neutral Policy cannot contain an absolute right of this sort, since some rescues would be prohibitively expensive to society—imagine someone who needs a hundred-million-dollar medical operation simply to prolong his life for one week. However, the Neutral Policy might well contain a more limited positive right such as "a right to be rescued *when* the net cost of rescue is less than the net benefit". There are other complications as well: ideally someone who imposes relatively few insurance costs on society—e.g. by living a relatively safe lifestyle, thereby minimizing his chance of *needing* rescue; or by agreeing to waive his future right to be rescued—should be allowed to direct the saved costs to his own moral values. But I will not discuss such details further here, since they are mostly just a matter of finding the maximally efficient solution; there is little to say about them from the perspective of theory-neutrality that has not already been observed by theorists with other perspectives.

The above applies to people with acute, extreme needs: someone who has come down with a life-threatening medical condition, say, or someone who is being mugged. However, it also applies to everyday needs: food, shelter, and so on. This allows me to say a little bit about distribution of unowned resources, a topic which I avoided earlier. A person who is so poor as to be starving to death is not in a very good position to pursue his values. At best he is ripe for exploitation by others, which represents the kind of concentration of wealth and power which I have argued we should avoid. At worst he will die, losing all ability to pursue his values. So when choosing a policy for distributing land, minerals, abandoned property, etc., we will probably want one which allocates to as many people as possible enough resources to keep themselves alive and spend at least some time pursuing their values.

What if there are unowned resources left over after everyone has been allocated enough to live on? We want them to be used efficiently, so they should probably be allocated to *someone*, who can then auction them off to whoever can make best use of them. Most of what I have said in the past few sections will be neutral toward the question of who should receive that initial allocation. However, the discussion in Section 3.4.1 is applicable here. We want resources to be directed to many possibly-morally-significant goals rather than just a few such goals, so want to avoid concentrating resources in any single group's hands. So something along the lines of "assign control of new resources to whoever has enjoyed the fewest resources thus far"—i.e. a broadly maximin distribution principle—is reasonable. We can assign the extra resources to people who are disadvantaged in one way or another: e.g. ones with physical disabilities, ones born into a cultural setting that causes hardship, or ones who have suffered some other sort of bad luck through no fault of their own. However, remember that the Neutral Policy is to try to maximize overall goal-fulfillment, it is committed to the maximin principle only insofar as that principle is favored by considerations of overall efficiency. If a situation arises in which, unusually, the disadvantaged are *not* the ones whose goals would most benefit from additional resources, the Neutral Policy will favor giving control of the resources to those whose goals *will* benefit most instead.

To summarize the chapter: we have theory-neutral reasons to adopt, and encourage others to adopt, a "Neutral Policy" with the following features. When morally uncertain, the Neutral Policy would have us promote existing people's ability to fulfill their morally-motivated goals, and also to bring new people into existence within reason. When conflicts arise between different people's morally-motivated goals—including our own, if we are not so morally uncertain as to have none—the Neutral Policy will usually advocate a classical-liberal approach in which each person is allowed to spend his own time and effort pursuing his own moral values as best he can within the constraints of allowing others to do likewise, and will have a fair share of world resources to help him do that. "Fair" will mean something along the lines of "according with a maximin distribution, insofar as this can implemented efficiently". For situations in which this approach is not practical—e.g. situations in which *any* option would seriously damage *somebody*'s ability to pursue his goals, or situations in which letting everyone go his own way would be clearly inefficient in terms of the citizenry's total ability to recognize and pursue moral goals—the Neutral Policy will favor democratic decision-making.

Of course, saying "this is the policy to which we would have had most reason to commit before we knew what theory-based reasons were going to apply" is not the same as saying "this is the policy to which we should remain committed, given our information about which objective moral theories are more plausible than which others". Sufficiently-persuasive theory-based reasons could justify disrespecting the artificial "rights" I have advocated here, or even changing our political institutions to terminate their respect for those rights. However, I do want to note one puzzle: if the reasons in question are so persuasive, why would it be necessary to violate anyone's rights? Why not instead just *use* those persuasive reasons to persuade the people in question to waive their rights? So to justify divergence from the Neutral Policy described here, one *not only* needs good theory-based reasons to diverge from it, but *also* a good explanation for why other people are unconvinced by those reasons and need to have their judgment overruled. I will not speculate here about when—and whether—normal, fallible human beings genuinely find themselves in such an epistemic position, but I suspect that it is rare; usually in cases of disagreement, the hypothesis that the agent is the one who is mistaken will be difficult to justifiably eliminate.

Chapter Four – Concrete Applications

In this chapter, I will bring theory-neutral reasons to bear on concrete moral issues in two categories: reproductive issues, such as abortion, and issues concerning the interests of people not yet alive, such as environmental conservation. Hopefully this will be interesting in its own right, and will also help make the concept of theory-neutral reasons clearer and less abstract. Of course, I cannot *resolve* these issues, since I cannot take a stand on objective moral theories. All I can do is identify some theory-neutral reasons which should be taken into account when dealing with these issues, and discuss how the Neutral Policy—which, the reader will remember, is the policy which, if we were making a choice in the absence of theory-based reasons, is the one to which we ought to commit—would have us approach them.

4.1 – Abortion and Other Reproductive Issues

I shall start with the issue of abortion. What should the Neutral Policy hold with respect to the question of "under what circumstances may pregnancies be terminated?" The first thing I should note, however, is that it is a mistake to look at the issue quite that narrowly. What is the difference between aborting a fetus or using birth control to kill gametes and prevent conception? What is the difference between killing a fetus before birth and killing an infant after birth? Of course the objective moral theory may draw distinctions at one or both of these places, but the Neutral Policy does not; as far as the Neutral Policy is concerned, the only relevant transition in this neighborhood is when the child becomes capable of independent moral reasoning—which presumably occurs well after birth. So a single policy will be needed for birth control, abortion, *and* infanticide.

The correct question is not "when may a pregnancy be terminated?" but "when may a potential future moral agent's existence be prevented?".

That is not to say there will not be *some* age-dependent differences relevant to theory-neutral considerations. As part of the Theory-Neutral Reason to Promote Success, one should avoid wasting limited resources which others might find useful. Looking at the situation in cold, economic terms—pretty much the only terms we can look at it in, without making theory-based commitments—the older a fetus or infant is, the more resources have already gone into producing it. Even if the mother sincerely believes that bringing an additional person into the world will have negative consequences, she should still explore the possibility—especially if she is already past the earliest stages of the pregnancy—that someone else will want to adopt the baby. It might well be better according to *everyone's* values if she accepts payment to finish gestating it: the adoptive parents will get a child with less expense than if they had to start from scratch; the mother will get payment which she can use to advance *her* values, and the world population size will not be any different than it would have been if the adoptive parents *had* started a new fetus from scratch rather than adopting the existing one.

4.1.1 – Ideal Population Size

Using the principles from Chapter 3, we can try to identify the optimal population size of a given region, from the perspective of theory-neutrality. Let us start with the following notion of *underpopulation* and *overpopulation*, based on the notion of goal fulfillment developed in Section 3.1:

> Call the world, or a region thereof, *underpopulated* if the presence of an additional person would increase the population's overall ability to control the course of events.
>
> Call the world, or a region thereof, *overpopulated* if the presence of an additional person would decrease the population's overall ability to control the course of events.

The obvious way in which adding another person could increase the population's overall ability to control the course of events would be by increasing the availability of labor. With an additional worker in the pool, it would be marginally easier to hire someone to do whatever tasks one expected to have good consequences. An additional individual's participation in the economy can also potentially allow for greater division of labor, thereby increasing the efficiency of *everybody's* labor—not only will it be easier to hire someone to try to bring about whatever consequences one wants to bring about, it will be easier to hire a *specialist*. We also should not overlook the possibility that the additional person will make intellectual contributions to our civilization: scientific and technological advancements should eventually percolate out to everyone, no matter how big the overall population is; so the more people there are who can make such advancements, the more such advancements everyone will be able to make use of.

On the other hand, adding another person can also have negative effects. Keeping a person alive requires the use of limited resources: food, energy, and so on. The resources thus consumed will therefore *not* be available to be used for trying to control the course of events. Similarly, adding another person will increase the total burden on social systems such as education, law enforcement, etc.; the increased cost of administering these systems must also ultimately come out of the population's resources.

More people also means that more resources are likely to be wasted on zero-sum activities such as producing advertisements intended to get consumers to reject one product in favor of someone else's equivalent product.

My guess is that if we did this calculation for the United States, we would find that we are still somewhat overpopulated. On average, citizens are receiving enough wages for their labor, even after tax, to allow them to buy enough food, energy, and so on to sustain themselves and their children without outside assistance, which—assuming that our market is setting approximately-correct prices on everything—suggests that the positives of economic participation are still generally outweighing the negatives of resource consumption and service use; which means that once we add in "increased rate of intellectual progress" to the mix, the positives win out. But a skeptic would argue that our market is not functioning correctly, that it has gotten skewed by unsustainable consumption of natural resources that properly belong to future generations, or by incurring foreign debt or otherwise exploiting residents of other regions. In any case, this is not really a topic for armchair philosophy: I have described *how* to make the judgment, but actually making it would require collecting, examining, and weighing the relevant data. My guess is just a guess.

Even if we turn out to be a little bit *over*populated with respect to the above definition, the Neutral Policy might still favor increasing the population further. We saw in Section 3.2 that when making decisions that influence the number of moral agents, we should take into account not just the fulfillment of already-existing morally-motivated goals, but also the fulfillment of the morally-motivated goals which the new agents have the potential to form. If increased competition for resources slightly outweighs increased availability of labor, and so slightly decreases the existing population's ability to fulfill its goals, this effect might be counterbalanced by the fact that the new person's goals, if different from any existing goals, will *only* be pursued at all if the person is brought into existence. Thinking back to Section 3.4.1's argument against resource consumption, we can observe that the difference between having nobody working toward a given goal and having one person working toward it is potentially *much* larger than the difference between having a million relatively-wealthy people working toward a given goal and having a million slightly-poorer people working toward it. So if the world is slightly overpopulated, deciding to further overpopulate it may amount to a case of "greatly advance one set of goals, at the expense of slightly setting back many other sets"; and it may turn out that advancing the one set—bringing the new person into existence, so that he can pursue the consequences he thinks best—will win out. Of course this only works in cases of *very slight* overpopulation; in cases of *severe* overpopulation, even the Neutral Policy will oppose further reproduction.

4.1.2 – Who Should Make Reproductive Decisions?

My discussion so far has focused on theory-neutral reasons for and against bringing an extra moral agent into the world. They should definitely be taken into account when trying to make a decision such as "should I use birth control?" or "should I have an abortion?", weighed alongside the instructions offered by whatever particular moral theories we place special credence in. But in our society, the abortion debate often takes a slightly different form: the question frequently asked is "should we, as a society, *ban* abortion?" I do not think such a ban could be justified under the Neutral Policy.

If people feel that increasing the population would have positive consequences, there are many options available to them that do not involve using the force of law to coerce pregnant women into refraining from having abortions. They could engage in reproduction themselves. They could hire a surrogate. They could offer some inducement to the pregnant women in exchange for those women refraining from having abortions—an option which also works if their view is not that increasing the population is morally positive, but rather that preserving existing lives, even fetal lives, is morally positive. In all these cases, *they* would be the ones spending resources to do what *they* think right, which is as it should be under the system described in Section 3.4. In contrast, it seems that forcing a woman to carry an unwanted pregnancy to term, when she herself—after considering all the facets of the issue—feels that that is not the best choice, would be to make her suffer the burden of fulfilling an objective moral judgment that she herself does not share. In short, it would be an appropriation of *her* share of resources for the purpose of advancing someone *else's* morally-motivated goals, which is exactly what Section 3.4 argued against.

Given Section 3.4's argument that it is better that people with minority moral views be allowed to act on those views, rather than be forced to act on the others' moral views, the only way society could justify intervening in a woman's considered decision to have an abortion would be by somehow arguing that this decision represented an unfair appropriation of resources to the causes she favors—"unfair" in the sense that the resources should have been allocated to someone else's cause, according to the framework from Section 3.4. It is hard to see how such an argument would go. Whose resources are being appropriated? I shall briefly survey a few possibilities, but I do not think that any of them are convincing.

Argument 1: a woman who has an abortion unfairly denies a benefit to future generations, which would otherwise have an extra member to help pursue their moral values. This cannot be quite right. If someone gets pregnant and then has an abortion, this does no more harm to future moral views than if she never got pregnant in the first place. So a pregnant woman cannot have any *special* obligation to the future. The Neutral Policy might be able to justify a legal obligation on each generation to bring the next generation into existence, but only if the burden fell equally on everyone. The most natural way to impose that burden would be by taxing the childless and rewarding those who gestate and raise children, and adjusting the schedule of incentives until sufficient membership in the next generation was assured—not by imposing a ban on abortions which burdened the unintentionally-pregnant severely and everyone else hardly at all.

Argument 2: a woman who has an abortion unfairly denies a member to the followers of whichever *particular* moral view the fetus would have grown up to hold, causing that group to have fewer resources under their control than they otherwise would have had. This might be plausible if the mother were somehow foreseeing what possibly-value-tracking moral judgments her child would eventually make, and wanted to abort it *because* she disliked those particular judgments. But as remarked in Section 3.2, it is difficult to *foresee* judgments which track actual moral values, if one does not already know what the actual values are. I suppose we can imagine a case in which doctors find genetic markers for predisposition-to-utilitarianism and predisposition-to-egalitarianism, but do not know the biological mechanism by which those genes function and so cannot say which one, if either, results in clear-headed reasoning and which results in a bias. More realistically, we can imagine a rape victim thinking "maybe I was raped because this child's father had a bad moral view, and maybe that moral view will turn out to be hereditary, so I had better abort the fetus just in case". But the vast majority of abortions—even the vast majority of abortions of the products of rape—do not have this sort of rationale. In most cases, the woman who wants an abortion would have no reason to expect her child's moral views to be less congenial to her than a stranger's child's moral views, so we cannot plausibly accuse her of attempting to skew the moral views of the next generation. In these cases, this argument is not going to work as a reason to ban abortions. If there is an unintentional but still systematic skew occurring—suppose, say, that being pro-choice turned out to be partly heritable, and so was less well represented than it ought to be given the quality of arguments for it—we would again do better to try to correct it some other way, such as by paying compensation to the followers of the skewed-against moral view, rather than by placing a special burden on pregnant women.

Argument 3: a woman who has an abortion is unfairly disposing of a resource to which someone else—such as the father or other relatives—had a claim. This is not completely implausible; it is true that we do not want—for roughly the sort of reasons described in Section 3.4.1—to systematically give fertile women more control over the course of events than we give to men or to infertile women. But banning abortion, or even allowing fetuses' fathers to veto abortion, would go much too far in the other direction: it would place on fertile women the burden of sometimes having their bodies appropriated to someone else's purposes, without compensation. In practice, even with abortions legal, they already suffer the burden of having to hire an abortionist and suffer the medical risks of abortion when faced with an unwanted pregnancy; I suspect this roughly balances out whatever special advantages fertile women enjoy, leaving them with not particularly more power to influence the course of events than everyone else has. Taxing their ownership of a healthy womb as though it were some kind of windfall, let alone regulating their use of that womb, would be going too far.

In Section 3.4, I argued that the Neutral Policy gives us reason to establish a society which respects individual freedom. We have seen here that this applies to the freedom to make reproductive decisions as well, except perhaps in the special cases in which somebody is deliberately trying to skew future generations' moral judgments—e.g. cases like "I will have an abortion, being I predict that my fetus will be predisposed to such-and-such view and I do not approve of that view", or "I will reproduce as much as possible, in order to have children and indoctrinate them into such-and-such view". The considerations given in Chapter 3, especially Section 3.2, *do* favor the creation of new people, to a limited extent, but they are considerations for individuals to take into account; they are not justifications for criminal laws against abortion.

4.1.3 – Human Engineering and Enhancement

Not all reproductive decisions affect the total number of people who exist; some just affect *which* people exist, or what traits the people come to have. For example, consider a mother who discovers that her fetus has unwanted genetic or developmental traits—Down syndrome, say—and is tempted to abort it in order to immediately begin gestating a new one. The effect on total population of such a decision, especially if made early in the pregnancy and if she really is able to get pregnant again immediately, would be minimal: the main effect is just to replace a person who could have existed with a different person, albeit one who is a few months younger. Other reproductive decisions—e.g. pre-implantation genetic diagnosis and selection, genetic engineering, selection of donor gametes, etc.—will have a similar effect.

Indeed, for my purposes it is completely irrelevant whether the decision in question really involves replacing the child in question with a new one, or simply altering the child's traits. Theory-neutral reasons are concerned with humans only as beings capable of recognizing and pursuing valuable states of affairs. They will be concerned with an action which changes a human's goals, or his ability to pursue those goals, whether or not the action simultaneously changes his personal identity. But if the action leaves his goals and abilities alone, even while changing other aspects of his identity, theory-neutral reasons will not be concerned. So my discussion here will apply not just to selective abortion and genetic engineering, but also to issues such as hormone therapy and vaccination, and even to mundane parenting decisions such as what kinds of food and education to provide to one's children. Most of my comments will also apply to those forms of enhancement that adults can choose for themselves: performance-enhancing or mood-altering drugs, surgery, etc.

Another distinction irrelevant for present purposes is that between *treatment* and *enhancement*. Parents who abort a "disabled" child in order to produce a "normal" one and parents who abort a "normal" child in order to produce a "superior" one may have different standing with respect to the true objective moral theory, but—with a few small exceptions which shall be noted as the discussion proceeds—theory-neutral considerations are concerned more with *what* traits are being altered than with how the altered version compares with the rest of the population. The generally consequentialist flavor of theory-neutral reasons is appearing again here: the key question is not "what is it permissible to do to people?" but "what kind of people would it be good to have exist in the future?".

From a straightforward application of the discussion from Section 2.3, we can see that there is a theory-neutral reason to perform interventions which make people better at making accurate moral judgments: e.g. enhanced intelligence, creativity, and perceptiveness. Correspondingly, there is a theory-neutral reason *not* to perform interventions which reduce these traits. There is a theory-neutral reason to enhance people's moral motivation, willpower, and whatever else would make them more likely to act on their judgments of moral value, regardless of those judgments' content. And there is a reason to enhance people's ability to implement their judgments, e.g. by increasing problem-solving skills, physical fitness, or strength.

With this last category, however, a word of caution is needed: the question is not "will altering such-and-such child's traits improve *that child's* ability to control the course of events", but rather "will altering such-and-such child's traits improve the ability of *the population as a whole* to control the course of events"? So if a given trait—say, charisma or manipulativeness—confers *only* a competitive advantage, makes its bearer better at getting what he wants *only* at others' expense, then there will *not* be a theory-neutral reason in favor of enhancing that trait. The most relevant theory-neutral consideration will be Section 3.4.1's argument against allowing power to become concentrated: we might approve of competitive enhancements if performed on someone otherwise likely to be unfairly disadvantaged, and disapprove of them if performed on someone already enjoying unfair advantages.

There are some traits which we might be tempted to enhance which theory-neutral considerations will neither favor nor disfavor. For example, we might be tempted to raise our children to have cheerful personalities, so that they will enjoy life and will be pleasant for others to be around. Unless this has an effect on their willingness or competence at achieving moral goals, theory-neutral reasons will tend to neither approve nor disapprove of it. Any reason we have for such an intervention would be theory-dependent—e.g. based on our judgment that "happiness is good" is more plausible than "happiness is bad". Likewise, neutral reasons will be indifferent toward cosmetic changes—e.g. genetic selection for hair color, or giving a child a tattoo—except insofar as they may impact the future person's ability to do what he wants to do. I should add, however, that when evaluating a possible manipulation we must take into account not just the intended effect, but also possible side effects: genetic engineering, and to a lesser extent pharmaceutical intervention, is at present a novel and risky technology, and there will be some theory-neutral reasons to exercise caution—especially regarding unimportant, cosmetic manipulations—so long as this risk remains. We will want to develop the technology, to enhance the options of future people, so should not ban its use entirely; but we will want to discourage usage patterns which create a risk that large numbers of future people will be killed or disabled by mistakes.

I have said that theory-neutral reasons will favor enhancing people's intelligence and creativity, and will be indifferent toward many other personality manipulations. However, this is not quite true, since theory-neutral reasons *are* concerned somewhat with diversity. As we saw in Section 3.4.1, we want every plausible moral view to have *someone* thinking about it, looking for arguments for it, and performing the actions which most cost-effectively further it. Insofar as people's personalities affect their predisposition to accept various moral views, we should oppose manipulations which reduce the diversity of personalities in the population, and support manipulations which increase that diversity. For example, it might be desirable to have some gloomy people who will be attracted to gloomy theories and some cheerful people who will be attracted to cheerful theories, so that the true theory, whether it turns out to be cheerful or gloomy, will have at least *some* relatively-easy-to-recruit supporters.

This brings us to a manipulation which the Neutral Policy does oppose: trying to manipulate a child's future moral views. This might be done genetically—e.g. by finding a gene for valuing welfare over freedom, or valuing mercy over procedural justice, and selecting for it—or educationally, e.g. by indoctrinating the child into a particular view rather than letting him make up his own mind. This sort of thing could end up skewing the beliefs of the population, and lead to some views being popular not because there were good reasons to believe them but merely because the population had been manipulated into believing them. So it should be avoided. One can try to equip one's child with the tools necessary to form true moral beliefs; one should not try to manipulate one's child into sharing *one's own* moral beliefs regardless of whether they are true. On the other hand, just as the Neutral Policy disapproves of manipulating children into sharing the beliefs of others, it will *approve* of manipulations aimed at making children more independent-minded.

So those are the theory-neutral reasons which a person should take into account when making enhancement decisions. *Who* should make such decisions? What laws about it are appropriate? In the case of interventions for adults, such as mind-altering drugs, it should presumably be the individual in question—with, perhaps, some social safeguards to make sure that he really *approves* of what he is doing rather than is doing it out of addiction or desperation. For children, the decision should be made by their parents, or whoever's resources are being used to raise those children—others who think that a given decision is not the best way to prepare the next generation should either be trying to persuade the parents to change their behavior, or else should be producing their own children. For fetuses, the decision should be made by the mother, for the same reason that abortion-type decisions should be made by the mother—anything else would leave women with too little control over their own bodies. Safeguards for children and fetuses could also be appropriate, but should be aimed not at dictating the choices believed to be best for future society, but rather at ensuring that parents are genuinely trying to enhance their children's capabilities rather than just trying to create useful servants for their own projects.

4.2 – Environmental Conservation and Other Intertemporal Issues

A number of interesting issues, most notably but by no means limited to environmental ones, involve decisions about how to allocate resources across time. May we consume natural resources such as petroleum now, or should we save some of them for future generations? How strong an obligation do we have to avoid engaging in activities that cause long-lasting environmental degradation, e.g. in the form of global climate change? What kind of obligation do we have to preserve endangered species, archeological sites, and suchlike for future generations to study? These are all questions of how to allocate resources between present people and future ones; however, past people may also get into the act. For example, should people's property and contractual rights continue to be respected after they die—e.g. someone who buys a gravesite for himself and demands that it be left eternally undisturbed—even if this restricts the opportunities of future generations? Should we really allow Constitutional amendments passed by previous supermajorities to constrain what a majority of the present Congress wants to do? And so on. This section will attempt to apply theory-neutral considerations to all of these intertemporal issues.

To begin with the obvious: we have a theory-neutral reason to try to increase moral agents' access to resources, regardless of when those moral agents are living. With respect to the past, if we know what dead people's wishes were, we should—all else equal—try to fulfill those wishes rather than deliberately defying them. With respect to the future, we should—all else equal—try to bequeath as much wealth to future generations as possible, including avoidance of unnecessary damage to the environment or depletion natural resources. This latter claim is true, incidentally, notwithstanding Parfit's famous observation that actions affecting future prosperity can also affect the identity of future people and so may not be bad for any individual person.[69] So long as the future agents with plentiful resources are more likely to fulfill true values than future agents with depleted resources, we have a theory-neutral reason to bring into existence the former rather than the latter, regardless of whether the two possible sets of future agents share any identity.

Of course, these "all else equal" claims are not terribly interesting. I doubt that anybody would seriously champion *unnecessary* defiance or *unnecessary* waste. The interesting questions arise in the case of conflicts. Suppose that we do not want to violate the wishes of the dead out of mere perversity, but rather because our idea of the good differs from the dead people's vision? Suppose that we do not want to deplete natural resources simply for the thrill of destruction, but instead want to *use* them for things which we believe, by appeal to theory-based reasons, to be morally good? How do we weigh our theory-based reasons to use the resources in question with our theory-neutral reasons to leave control of those resources to others? These are questions for the Neutral Policy.

Assuming we can treat as negligible the possibility mentioned in Section 2.4.2 that morality will turn out to be culture- or time- relative, the Neutral Policy will call for us to be evenhanded between people living at different times. It does *not* endorse the kind of time discounting of the form "fulfilling a hundred goals today is as valuable as fulfilling a hundred and five goals—ones of comparable importance to the people who possess them—next year" which economic planners might be inclined to accept. Fulfilling a hundred of this year's goals is worth no less and no more than fulfilling a hundred of last year's goals or a hundred of next year's goals, all else equal.

However, we shall see that all else is *not* equal. When allocating resources, the Neutral Policy, while not concerned with temporal position as such, *is* concerned with factors related to temporal position: factors such as whose goals are most likely to be motivated by true moral beliefs, who can make the best use of a given limited resource, and who enjoys the most advantageous distribution of those resources which stand outside of our control.

Let us start with the question of whose goals are most likely to be motivated by true moral beliefs. First should be a reminder: as noted in Section 3.1, only humans—or, at most, humans and a handful of other animals—are capable of forming and acting upon moral beliefs. Creatures like beetles and trees are presumably incapable of moral thought. Theory-neutral considerations are therefore not concerned with protecting their interests, any more than they are concerned with other environmentalist goals like preserving nature or beauty; environmentalism, at least when it goes beyond simple conservationism, can only be justified by theory-based considerations. Anyhow, moving on from beetles and trees to beings which *might* be capable of moral thought, such as great apes or cetaceans: even those beings have such limited intelligence that their ability to influence the course of events will be extreme low no matter how we behave. So even if they end up going extinct as a result of anthropogenic climate change, I suspect that their degree of moral-goal-fulfillment will not have been reduced significantly. Since this is a section about weighing interests of beings at different times, I suppose I should acknowledge that there may one day be non-humans—artificially-intelligent robots, or genetically-engineered post-humans, or "uplifted" animals, or something—which *are* as capable, or even *more* capable, of effective morally-motivated action as we are, in which case they will matter as much or more than we do. But the point remains that, at present, the Neutral Policy can safely ignore effects on *wildlife* when considering environmental issues.

What about future people? Here, I think we should be very concerned. Future people are *more* likely than we are to have accurate moral beliefs. My first argument for that claim is not theory-neutral, but worth making anyway: the general trend of human history, at least since the invention of the printing press, appears to have been positive. We have made significant moral progress over time, gradually recognizing slavery, racism, sexism, and so on as morally abhorrent, and generally expanding our sense of human rights. That is not to say that nothing has been forgotten—perhaps we have lost sight somewhat of important values like loyalty and communal solidarity—but I think the losses are far smaller than the gains. Skeptics might be inclined to note that many of the worst manmade disasters of human history took place during the twentieth century, after alleged centuries of moral progress; but I think this is a mistake. I suspect that if Nero or Attila had possessed armies numbered in the hundreds of millions, and weapons such as poison gas and atomic bombs, they would made modern-day tyrants such as Hitler and Stalin look positively gentle. The increase in manmade disasters over time is thus better explained by an increase in man's ability to cause disasters, not by a worsening of human character.

Returning to the theory-neutral spirit of this dissertation: I think one can make a compelling case that we should expect moral progress over time, *without* appealing to any particular instances of alleged progress. Once it becomes sufficiently easy to preserve and disseminate intellectual discoveries—including discoveries in the field of ethics, or fields which inform ethics—each thinker can stand on the shoulders of those who came before him, and see further. Likewise for the related field of character education: hopefully, as time passes, we will not only become more knowledgeable about what is right, but also better at motivating people to *do* what is right. In a world with printing presses, let alone a world with the Internet, information *will* tend to accumulate over time; how could this accumulation not, in the long run, lead to progress in moral knowledge and moral motivation?

So I believe that the average person alive today is more likely to be pursuing actually-good consequences than the average person who lived a hundred years ago was, and *much* more likely to be pursuing them than the average person who lived a thousand years ago. I also believe that, barring the utter collapse of civilization or the rise of some sort of absolute dictatorship concerned only with self-preservation, the average person alive today is *less* likely to be pursuing actually-good consequences than the average person alive a hundred years from now will be, and *much* less likely than the average person alive a thousand years from now. This gives us a reason to preserve resources for future generations rather than using them ourselves. If we use them today, we may waste them on goals which turn out to be bad or chimerical; if we save them for the future, they will more likely be used for morally important goals. Of course, the force of this reason depends on how far along we are in finding the true moral theory. Even if moral progress were the sole consideration—and we shall see momentarily that it is not—it would be ridiculous for me to advocate conserving resources *forever* and never using them; eventually, when we can reasonably believe that *most* moral progress has already been made and that all that remains is some minor fine-tuning, it will be time to implement our moral beliefs. But for now, I think there is much progress remaining to be made, and this counts in favor of allocating resources to future people rather than present ones—and to present people rather than past ones, and to moderately-distant-future people rather than immediate-future ones, and so on.

People's likelihood of having accurate moral beliefs is just one part of the picture. Another important element in deciding who should have access to a given resource is how useful the resource would be to that person, how much it would help that person fulfill his goals. One might at first be inclined to think that this also favors future people. After all, just as they will know more about ethics and character-building than we do, they will also know more about how to efficiently consume resources. Consider how early consumers of petroleum used to burn natural gas at the drilling site rather than capturing it and using it like we do today: we would have gotten more total energy out of their wells than they got.

However, I think increases in efficiency over time are completely outweighed by a factor that favors the consumption of resources by relatively early people: because the future is malleable and the past is not, being positioned early in history gives one more power to influence it. A verse from the rock opera *Jesus Christ Superstar* comes to mind here:

> You [Jesus]'d have managed better if You'd had it planned
> Now why'd You choose such a backward time and such a strange land?
> If You'd come today You could have reached a whole nation
> Israel in 4 B.C. had no mass communication![70]

Think about that for a moment. If Jesus's message had originated two thousand years later than it actually did, would it have reached more people? Certainly not. First, none of the people who lived during those intervening two thousand years would have gotten the message. Second, "backward" means of communication are not *so* slow that they were unable to reach every corner of the Earth when given two thousand years in which to operate. Third, once mass communication *was* finally invented, the hundreds of millions of people who had received the message by "backward" means were all free to start spreading it with the new technology as well. The lesson here is that the earlier one acts, the more ramifications one's action can have across history—even if, by virtue of acting early, one has chosen a "backward" or "inefficient" strategy. As soon as we know approximately what the right goals are—whether they are "spread such-and-such message to as many people as possible", "make as many people as possible as happy as possible", or what—we should start trying to implement those goals: unnecessary delays would significantly reduce our ability to influence the course of events.

Another consideration in the "who can use the resources best" category, also favoring present people, is the fact that we do not really know *what* resources future generations will find useful. The relevant proverb here is "people in the nineteenth century worried that the world was running out of whale oil". It would have been a mistake for such people to attempt to conserve whale oil for *us*; we already possess substitutes that we prefer. For stone-age people to try to conserve their flint mines for us would have been even more wasteful. They had a use for the flint; we do not. If we try to conserve resources for future generations, we may end up making similar mistakes—significantly decreasing our own ability to influence the course of events, while not significantly increasing *their* ability to influence it.

I am not sure which consideration—the future's better moral knowledge, or the present's better position to influence the course of history—is larger; to some extent they cancel each other out. A tiebreaker, perhaps, will be Section 3.4.1's argument against concentration of resources. Suppose that past, present, and future people all have plausible ideas about what is morally good—but that they have systematically *different* ideas. Of course, to the extent that this is not true and that future people will share our values, they will not object if we decide to borrow some of their share of resources so as to apply those resources earlier in history where they can have a larger effect; but I think we would be unwise to rely too heavily on this assumption. If we think that some degree of systematic disagreement across time is likely, then, from the argument in Section 3.4.1, we will want each generation to have access to some resources, so that each generation's highest priorities can be attained. Setting aside issues of moral progress and of efficient use of resources, if history is going to include a trillion total agents, then ideally each agent would get to use one trillionth of the world's limited resources. Of course, this ideal is significantly complicated by the fact that we have no real clue what the total population of moral agents across history will be. A hundred billion? A trillion? Ten trillion? More? It is also complicated by the whale-oil-and-flint-mines problem: it is hard to say exactly what counts as a limited resource. If future people will use "resources" that we currently are not even touching, that entitles us to a larger share of the resources we *are* consuming. To give one simple example: we are not presently consuming *any* of the minerals deposited in the asteroid belt; does that entitle us to consume a relatively large share of the minerals deposited on Earth? How about the many materials on Earth that are currently regarded as essentially worthless, such as saltwater; perhaps future generations will be able to consume them, and so will not begrudge us our—e.g.—petroleum usage.

To some extent these complications also cancel each other out. The longer human civilization ultimately lasts, the larger the total population across time will be; but the longer it lasts, the more previously-useless or -inaccessible objects will ultimately count as someone's share of resources. Quite possibly the easiest-to-implement solution of "each generation is allocated the right to consume whatever resources it can acquire and use" will not stray *too* far from the ideal of equal allocation. Indeed, it may plausibly end up allocating *more* resources to future people than to present, once one includes the manmade resources we will be bequeathing to future generations: most notably intellectual advances in fields such as medicine and technology, which are likely to prove extremely useful to future generations, but also feats of engineering—e.g. roads and canals—which last multiple generations, and even consumer goods whose raw materials are easier to reclaim via recycling than they were to mine initially.

If we want to give even more influence to early-in-time people than the "resources may be used by whichever generation first wants them and reaches them" approach, we can do so by manipulating legal rights. In addition to letting people *consume* resources, we also currently let them acquire title to *future* consumption of certain resources. Most notably: in our society, if I buy land, I acquire the legal right not only to build on it, mine minerals—all minerals, even those I presently have no interest in mining—from it, capture the sunlight that hits it, and so on while I am alive, and *also* the right to bequeath the right to do these things to the entity of my choice after I die. If I own enough land, I can in principle set up a trust fund which maintains itself with some of the profits of temporarily renting out usage rights to the land, and which uses the rest of those profits to further *my* agenda even centuries after I am dead. Things like this happen sometimes, and can cause messes when morals change over time. For example, there was once a case in which a very rich racist, in a time when racism was widely viewed as reasonable, chartered a university, with the stipulation that only white students be admitted; sixty years or so after his death, it had become clear that racism was morally wrong; the university wanted to change its admissions policy, but was bound by the terms of the charter.[71] Happily the trustees discovered a loophole, convincing a jury that the founder's intention had been to create a "first-rate" whites-only university, that this had become a contradiction in terms, and that dropping the "whites-only" part was less damaging to the founder's intention than dropping the "first-rate" part; but such loopholes will not always be available.

I suspect that systems of property rights which allow trusts and endowments and the like to endure forever, and even grow over time, go too far in favor of people early-in-history. Consider: there is—I do not say "suppose there is", since this is another true story, albeit a more generic one—a starving and impoverished man, wanting to consume his share of the world's natural resources; the natural resources in question are available, since past generations were unable to consume many of the particular ones he is interested in, such as "a plot of land in the twenty-first century, and the sunlight and rain that falls on it", which he could use to acquire primary goods like food and income; and yet he cannot consume those resources because he finds that they are all owned by other people—ownership claims which were established before he was even born, although some may have been transferred since then. He never had a chance to claim any natural resources; he would have to have been born sooner. Since his fair share of natural resources is surely more than zero, something has gone wrong. Plausibly we should require anyone who claims permanent title to *all* of the resources within a given section of the planet, rather than just the right to pick up and use whatever individual resources from that section he has an immediate use for, to pay a periodic property tax into the common coffers, to be distributed to present people as compensation for their share of the resources found in the claimed section. The property tax should be *larger* than the rent which is being collected on the property, since otherwise there would effectively be a net transfer of wealth from the present generation that is renting the property to the past generation that owns the estate.

The reader will probably be finding my whole picture to be rather Lockean.[72] People can use any resources with which they can "mix their labor", as it were. But if they try to claim so many resources that future comers find there are not "as many and as good" resources remaining, then fair payment is owed to those future comers. And in any event, their right does not extend as far as permitting them to *waste* resources. Given the familiarity of these ideas, it is probably time for one of those periodic reminders: I have not committed to a theory of natural rights, nor to *any* objective moral theory. I am advocating this position based on the theory-neutral arguments from the previous chapters. The idea of letting each person—including future people—have control over a share of resources comes mainly from Section 3.4, and the idea of avoiding waste comes mainly from Section 2.3.3. Maybe there are other, less Lockean, ways to implement those ideas; but I do not see how.

I have not yet said anything about one of the cases I mentioned at the start: Constitutional amendments. It makes sense to *have* a Constitution that cannot be overruled by a mere majority, to protect democracy and individual rights—which, as we saw in Chapter Three, are advocated by the Neutral Policy—from being swept away by tyrannies of the majority; and no doubt it will need amendments from time to time to help it serve this purpose. But the amendments process should not be usable for any purpose whatsoever. Consider, for example, the Eighteenth Amendment to the Constitution of the United States, which imposed a nationwide ban on alcohol production and sales. Why did that need to be an amendment? Of course, the United States is at least theoretically a government of limited powers, so perhaps it was necessary to pass an amendment granting Congress the *power* to ban alcohol and to enforce such a ban—although it is not immediately clear to me that a government *should* have the power to restrict individual liberty in such a way, and anyhow the lack of such an amendment does not seem to have hindered the American federal government's attempts to control other psychoactive drugs—but why was it necessary to include the ban itself as part of the amendment? It seems to me that the 1917 Congress, which passed the Eighteenth Amendment, was simply trying to exert control over the behavior not just of American citizens of the 1910's—who were the people who had elected it to represent them, and so the people for whom it *should* have been making laws—but also over American citizens of later decades. This was rather hubristic of it: those later citizens were foreseeably going to know more about the effects of an alcohol ban; they were foreseeably going to know more about biochemistry and sociology and the various other fields which would inform alcohol policy; and they were foreseeably going to have more advanced moral theories and political philosophies. For that matter, they were also foreseeably going to be more numerous. In 1917, the American population was about 100 million; achieving a two-thirds supermajority for an alcohol ban required the ban to be supported by about 33 million more people, or rather their representatives, than opposed it. By 1933, when the ban was repealed, the population was over 120 million; getting a supermajority required a margin of over 40 million Americans' representatives. So the average American voter from the 1930's had effectively *less* say over 1930's American alcohol policy than the average American voter from the 1910's did, despite being better informed about the issue. Not to mention the asymmetry involved: 1930's Americans had *no say whatsoever* about 1910's American alcohol policy. Something seems to have gone wrong with the distribution of political power between those generations. I think the moral of this story is that a Constitution—or rather, anything requiring more than a simple majority to alter—should be used only for establishing the extension and limits of government power, and the basic structure of government; *not* for trying to etch particular policies into stone.

In this section I have described the various sorts of considerations that need to be balanced when dealing with issues involving future generations, and given some examples of such issues. If we are uncertain about what is right for us to do, we should try to leave future generations—who will hopefully be better informed than us—as many resources as possible. More generally, we should try to leave them as large a range of *options* as possible, whether these are consumption options or legal options or what-have-you. Even if we think we do know what is right, we should still refrain from *unnecessarily* consuming resources or constraining future options; we may consume resources, pass laws, and so on in pursuit of our values, but should not try to claim power over future generations' consumption patterns and other choices.

Chapter Five – Conclusion

I have argued in this dissertation that some types of action have the feature of being approved, or at least being relatively likely to be approved, by a very large variety of objective moral theories. As long as we put any credence at all in some of those theories, even if we put no more credence in them than in their opposites, this fact means that we can subjectively expect the actions in question to have, all else equal, a higher-than-average probability of being morally right, and of being morally right to a higher-than-average degree. My argument has been theory-neutral: I have neither assumed nor argued that particular theories which approve of the actions in question are any more plausible than particular theories which disapprove of them; rather, what I have argued is that theories which approve of them are far more *numerous* than theories which disapprove of them. Noticing this fact about an action—noticing that it is much more widely approved than its alternatives—will generally tend to raise, and will never lower, our estimate of its subjective rightness, no matter what our credence distribution over moral theories looks like. So I have called it a "theory-neutral reason" to perform such actions. Theory-neutral reasons will be relevant to almost everyone's decision-making, even people who are facing severe moral uncertainty or are otherwise unable to appeal to theory-based reasons.

Chapter Two identified three main types of action we have theory-neutral reason to perform: ones which are expected to increase the accuracy of people's judgments about what consequences ought morally to be pursued, to increase the motivation of people to pursue the consequences they have judged to fall into this category, and to increase the likelihood that people so motivated will succeed in bringing about the consequences they are pursuing. If any at-least-partly consequentialist moral theory—that is, any theory which would assent to the judgment "there are some consequences such that anybody who brings about one of those consequences, whether directly or indirectly, will tend to have acted morally better, all else equal, than if he had not brought it about", regardless of whether it prefers high-utility consequences, high-justice consequences, or whatever, and regardless of whether it places deontological side-constraints upon our pursuit of the morally good consequences—turns out to be true, then such actions will have an above-average likelihood of indirectly bringing about the consequences which are morally right to bring about, and therefore an above-average likelihood of being morally right. So we have theory-neutral reasons corresponding to these three types of actions. I also identified a few other theory-neutral reasons which do not quite fit into this schema: a theory-neutral reason to imitate others even when they are engaged in non-consequence-directed activity, and a theory-neutral reason to try *especially* hard to improve the judgment, moral motivation, and success of people who are relatively similar to ourselves.

Chapter Three discussed what I called the "Neutral Policy"—norms and rules which we have theory-neutral reasons to adopt. The Neutral Policy includes, on an individual level, vaguely-utilitarian-looking recommendations aimed at increasing the fulfillment of people's non-self-centered goals. I suggested that we try to fulfill potential goals whose existence is contingent on our behavior—that way there will be no temptation to give ourselves extra, undeserved, credit for bringing into existence a goal that was automatically fulfilled, nor for removing from existence a goal which was not fulfilled. I also argued that the Neutral Policy will favor liberal social institutions which allow individuals to pursue their own goals as much as possible—as a way to ensure that all plausibly-valuable goals receive some pursuit, and because it is inefficient to try to set people to tasks which they do not support—and which handle any decisions which must be made collectively in a roughly democratic manner. The burden of proof, in ethics and political philosophy, is therefore on people who are *against* utilitarian personal behavior and liberal institutions. Until and unless they can present a sufficiently-persuasive argument for particular moral theories which oppose such policies—i.e. can raise their subjective credence to higher than the combined credence in the relatively large number of moral theories which support those policies—promoting them will be the action with the highest subjective likelihood of being morally right.

Chapter Four gave some concrete examples of how the theory-neutral reasons and the Neutral Policy might be applied. Regarding reproductive issues such as abortion, people have—provided that the world is not *too* overpopulated—a theory-neutral reason to have children, so should only use abortion or other forms of birth control when they believe that childbearing would have morally bad consequences overall, e.g. due to reducing their own productivity or due to some negative value attaching to the child's life; however, if a given individual *does* judge that refraining from childbearing will improve her ability to function, other people should refrain from interfering with their decision. Regarding human genetic engineering and other enhancements, theory-neutral reasons and the Neutral Policy will support ones that make people more effective at recognizing, wanting to pursue, and pursuing morally good consequences, and will be neutral toward ones which do not affect people's effectiveness at these activities. Regarding our relationship to past people and future ones: theory-neutral reasons maintain that we should respect people in both categories, just as much as we would respect presently-living people of similar character and informedness; so the Neutral

Policy will be to respect dead people's property and contractual rights—while making sure that people cannot acquire so many such rights as to leave the living with no opportunity to acquire property of their own—and to try to leave for future people as many resources as we can, and at least as many as we ourselves have access to.

Of course, it is important not to lose sight of the fact that my discussion *only* covered our theory-neutral reasons, not our theory-based ones. The claim that theory-neutral reasons or the Neutral Policy favor a particular action is not at all in conflict with the claim that a given moral theory disfavors that action, nor with the claim that such a moral theory is the objectively true one. Sometimes a theory-based reason will conflict with a theory-neutral reason. This does not give us reason to think that the argument for either was mistaken; it merely forces us to weigh the two reasons against one another, comparing their strength with one another, and also comparing our credence in the one particular theory underpinning the theory-based reason with our credence in the disjunction of the many various theories underpinning the theory-neutral reason. I will say a little bit more about such comparisons in Section 5.1.

Also, my discussion is not the final word in any of these matters. There are various points of my argument, whether for the theory-neutral reasons paradigm in general or for the way in which it should be applied to some particular issue, where a reader might well disagree. Section 5.2 will discuss a few of the possible divergences from my line of argument—points at which I made one assumption about meta-ethics or empirical facts or whatever, and at which an alternative assumption would have led to different results—which I think have the most potential to lead to interesting avenues of thought.

5.1 – Quantification of Theory-Neutral Reasons

Some readers will still be bothered by the question of how to weigh theory-neutral reasons against theory-based ones. There are limits to what I can say about this question because identifying theory-based reasons and determining their strength is far outside the scope of this dissertation: it is not my place here to look at particular arguments for particular moral theories and speculate on how a reasonable agent would respond to those arguments. However, what I can do is give a few hypothetical scenarios of how a person who *did* evaluate the available theory-based considerations in a particular way would then be entitled to weigh them against theory-neutral reasons.

This discussion will be explicitly quantitative. Readers who prefer the more qualitative tone which the rest of this dissertation has taken may wish to skip this section; I include it only for the people who will not be satisfied until they have been shown some precise, fully-specified, numerical examples. I will be imagining here that people think in numerical terms: that they can say things such as "oh, yes, I have a 15.3% credence in utilitarianism, and that's a recognizably different mental state from the one I would have if I had a 15.2% credence in it or a 15.4% credence in it". I realize that nobody is actually that precise; we actually think in fuzzy terms such as "not impossible, but not probable either". Nevertheless, it is a useful idealization for purposes of discussion, since the reader and I are much less likely to disagree about what a "10% credence" is than about what "not probable" means.

To quantify theory-neutral reasoning, we need to select a specific strategy for decision-making under normative uncertainty: which strategy we choose will affect the details, if not the broad strokes, of how to weigh different kinds of subjective reasons against one another. For present purposes I shall use the "subjective expected rightness" measure described in Section 1.1.1, which says we should choose the action whose average degree of moral rightness across different hypotheses about the objective moral truth, weighted by those hypotheses' subjective probability, is highest. Readers who prefer some other strategy, such as "minimize the subjective probability that your action is morally awful", will have to make the necessary changes when reading my discussion; I will say more about what changes would be necessary in Section 5.2.

Let us begin with a very simple example to demonstrate how the expected rightness calculus is meant to work. Suppose that there are two equivalence classes of possible outcomes—that is, sets of outcomes such that the members of any given set are identical in all morally-relevant respects—of a decision: call them X and X^C. Suppose that, on the basis of theory-based reasons involving the various moral theories which we have judged to be plausible, we find ourselves extremely uncertain which is better, but leaning ever-so-slightly toward X: we have a 55% subjective confidence that actions resulting in an outcome from X have a moral rightness of 1 and actions resulting in an outcome from X^C have a moral rightness of 0, and a 45% subjective confidence that the situation is reversed, that actions resulting in an outcome from X^C have a moral rightness of 1 while actions resulting in an outcome from X have a moral rightness of 0. Suppose that we are choosing between three options: we can directly bring about an outcome from X, we can directly bring about an outcome from X^C, or we can pass the decision off to some other agent who will inform himself about the issue and then make a selection. Suppose that we believe him to be slightly biased *against* X: we estimate that if bringing about an outcome from X is the right option, he has only a 40% chance of correctly choosing that option and a 60% chance of wrongly choosing to bring about an outcome from X^C, whereas if bringing about an outcome from X^C is the right option then he has an

80% chance of correctly choosing X^C and only a 20% chance of choosing X. Given all this, should we go with our theory-based judgment that X is most likely to be best, or should we go with the general theory-neutral reason to empower others to do what *they* think best?

In this scenario, the expected rightness of choosing X directly is straightforward: there is a 55% chance that it has value 1 and a 45% chance that it has value 0, so its expected rightness is 0.55. Similarly the expected rightness of choosing X^C directly is 0.45. What is the expected rightness of passing the buck? Well, the chance that choosing X is the right option and the agent will choose X is 55%*40% = 22%. The chance that X^C is the right option and the agent will choose X^C is 45%*80% = 36%. So—notwithstanding his slight bias in favor of a consequence which we suspect to be the wrong one to pursue—the overall chance of the agent making the right choice is 58%; since choosing rightly has value 1 and choosing wrongly has value 0, the overall expected rightness of enabling him to make the choice is 0.58. So in this particular case the theory-neutral reason to empower others narrowly scrapes past the theory-based reasons in favor of bringing about an outcome from X, and the subjectively best choice is to empower the other agent to control the outcome. But if we had been slightly more confident in our own moral judgment, or slightly less confident in the other agent's judgment, the calculation could have gone the other way and the subjectively right option would have been committing to X.

In practice, of course, applying the expected value calculus will be much more complicated than this. Normally one's options are not "choose from this set of options or empower someone else to choose between them"; instead they are "choose from this set of options or empower someone else to choose from a different set". Indeed, the two sets of options cannot be precisely identical, since the world will have a different history depending on who made the choice, and this different history could in principle affect the moral status of the outcome—that is why I said "equivalence classes of outcomes" above, rather than just "outcomes". But it is unlikely that the history will be the only difference. For example, consider a case of charitable aid: suppose an agent from a reasonably wealthy country is deciding whether to spend ten dollars on a meal at a local restaurant, spend it on a book from a local bookstore, or send it to a poor person overseas. Certainly the aid recipient will not be deciding between that restaurant and that bookstore! Often the agent will not even know what options the aid recipient will face, but may have divided credences over hundreds or even thousands of possibilities. How can the agent weigh the theory-neutral value of empowering someone when he knows next to nothing about that person's situation?

Of course the agent can still do the expected rightness calculus. He could consider all the many different things he can imagine the aid recipient doing with the money, and the many possible values each of those things could hold; and then somehow assign a probability to each and every hypothesis. However, the problem quickly becomes unwieldy. An alternative is to try to estimate in abstract terms the expected usefulness the two individuals would make out of the money when faced with an arbitrary situation. This would involve asking four questions. First, which of the two is more likely to make accurate moral judgments, and to what extent? Perhaps the agent is more likely to recognize moral truths, if he is better educated; perhaps the aid recipient is more likely to recognize them, if he has more experience with morally-significant situations. Second, which is more likely to act on his moral judgments, and to what extent? Perhaps the agent is more likely to be morally motivated, since he can afford the luxury of altruism; or perhaps the aid recipient is more likely to be morally motivated, if his tougher life has made him less soft and spoiled than the agent. Third, which one is more likely to use the money effectively, and to what extent? Perhaps the agent will get more marginal goal-success from it, again because of his education, or perhaps the recipient will get more success, if he lives in a country where the money can go further or is closer to the brink of failure. Fourth, which one is more likely to be involved in morally-significant situations, and to what extent? Perhaps the agent is more likely to face important situations, if he has more information sources and so can concern himself with events further away; or perhaps the aid recipient, if such situations are a more common part of his everyday life. The agent must estimate as best he can the answers to all of these questions. Even these may seem like a lot of different questions to juggle, but asking four questions about each of two people is much simpler than considering thousands of different combinations of hypotheses about which moral theory is true and about what a poor person might believe.

Now let me spell out how to use the answers to those general questions. Consider a future decision being made by some individual—either the agent or the potential aid-recipient—between outcome-classes X and X^C. Let Gx be the claim "X is morally better than X^C", Jx be "the individual judges X to be morally better than X^C", Mx be "the individual will try to make X occur instead of X^C", and Rx be "X occurs instead of X^C". Let ~Gx be "X^C is morally better than X", ~Jx be "the individual judges X^C to be morally better than X", and so on—not worrying too much about cases where X and X^C are equally good, or the individual judges them to be equally good, or the agent withholds judgment, or whatever. Now we can estimate the likelihood of this particular individual bringing about morally good consequences. Define a, b, and c such that

$P(Jx|Gx)=P(Jx|{\sim}Gx)+a$, $P(Mx|Jx)=P(Mx|{\sim}Jx)+b$, and $P(Rx|Mx)=P(Rx|{\sim}Mx)+c$—I use here the familiar symbolism of decision theory in which $P(Q|R)$ means "the probability of Q given R" and is equal to $P(Q\&R)/P(R)$, which is to say "the probability of Q and R both being true, divided by the probability of R being true". In English, those definitions amount to: a is how much more likely the individual is to judge that X is good if X really is good than if it is not, b is how much more likely he is to pursue it if he judges that it is good than if he does not, and c is how much more likely X is to occur if the individual pursues it than if he does not. An individual who is a perfect judge of moral value has an a of 1, and an individual whose moral judgments are completely arbitrary has an a of 0; an individual who is motivated solely by moral concerns has a b of 1, and an individual who is motivated solely by non-moral concerns has a b of 0; and finally, an individual who has full control over what happens has a c of 1, whereas an individual with no control has a c of 0. We also need to assume a few independence claims: $P(Mx|Jx\&Gx) = P(Mx|Jx)$ and $P(Rx|Mx\&Jx\&Gx) = P(Rx|Mx\&Gx) = P(Rx|Mx\&Jx) = P(Rx|Mx)$. That is, whether the individual tries to achieve X does not depend on whether X is good except insofar as X's goodness might influence whether the agent judges X to be good; and whether the individual does achieve X does not depend on X's goodness or on the individual's judgment about X's goodness, except insofar as these might influence whether the individual is motivated to pursue X. Finally, assume that $P(Gx)=0.5$, which will be true if we do not know what the individual's future decision is going to be about. Given all this information, the likelihood of the better outcome of X and X^C being reached—that is, of $P(Rx\&Gx)+P({\sim}Rx\&{\sim}Gx)$—turns out to be $0.5+abc$.[73] If v is the estimated degree of moral value for reaching the better of two possible outcomes in an average decision faced by this individual, then the value of his decision is $0.5v+abcv$. If letting this individual have extra funds will raise his effectiveness at achieving his goals from c to c', then the value of letting him have those funds is abv*(c'-c). So if we can estimate the value of a, b, c, c', and v for an arbitrary decision by the agent, and also estimate them for an arbitrary decision by the potential aid recipient, we can then figure out which one can make the most use out of the money.

In English: the time when we should most concern ourselves with promoting others' ability to do morally good things—which is at the heart of theory-neutral reasons and Neutral Policy—is when we have high faith in others' competence to make moral judgment and in their good intentions, when we can make a significant difference to their ability to achieve their goals, and when we expect them to find themselves in morally-charged situations. We should go with our own value judgments especially in cases in which we estimate that our own ability to make such judgments is significantly above-average, in which we think that the people we could spend effort trying to empower are unmotivated by considerations of moral value, in which those people's success or failure is not very dependent on whether we help them, and in which those people are relatively unlikely to find themselves in morally-significant situations. Of course, "empower others" and "do as you yourself think best" do not necessarily conflict with one another, since both may well involve beneficence, promotion of liberty, etc.—the various actions which are advocated both by theory-neutral reasons and by many of our leading objective moral theories.

It is probably worthwhile for me to include a warning here: given the natural human tendencies to exaggerate our own judgment abilities and to put a positive spin on our own unreasonable behavior, we must be very careful not to ignore theory-neutral reasons in situations where we should attend to them. "Everyone except me is a fool or a crook, but I am very wise and virtuous; so I will ignore their moral values and focus on fulfilling my own" may be *valid* reasoning—if it were true that everyone else were a fool and a crook, one would be justified in ignoring their values—but it is rarely *sound*. In fact, it seems more plausible for one to assume, unless in possession of convincing evidence to the contrary, that his contemporaries' ability and inclination to identify and engage in morally worthwhile activities are roughly the same magnitude as his own. If so, then he should aim to maximize total human success at goal-fulfillment—help others achieve their goals whenever such help advances their goals more than it hinders his own. In a situation where information is especially lacking, that is the strategy which will make the most sense—which is why the Neutral Policy came out the way it did in Chapter Three. Also, given the observed progress made in both ethics and technology over the past few centuries, it is not unreasonable to save for, or invest in, the future, on the grounds that future people will make better use of whatever we leave for them than present people—ourselves or others—can. Aside from those broad generalizations, I will not speculate further here on what our quantitative estimates of the strength of theory-neutral reasons ought to be.

5.2 – Other Directions

This section will discuss possible future directions in which one might try to expand theory-neutral ethics, and possible deviations from my discussion. The most obvious thing to do would simply be to look for theory-neutral reasons other than the ones I have discussed. This dissertation is a first foray into the world of theory-neutral reasons; I cannot claim that it represents a complete, meticulous exploration of that world. I have not attempted to argue that the theory-neutral reasons I have identified are exhaustive; and indeed we already saw, in Section 2.4, that not all theory-neutral reasons fit comfortably within Section 2.3's framework of "try to perform actions whose consequences are relatively likely to be worthwhile, due to the fact that somebody or other views them as worthwhile".

The other option, which I will be focusing on in this section, would be to change the assumptions on which my argument was predicated. Some such changes would completely undermine the notion of theory-neutral reasons; others would just affect the details of those reasons' content.

My most important assumption was the rejection of strong moral skepticism. Theory-neutral reasons are relevant to *weak* moral skeptics, people who accept that there might be discoverable objective moral truths but merely deny that they themselves have discovered those truths; indeed, theory-neutral reasons are at their strongest in the face of such skepticism, since if an agent has no information about what objective moral claims are true, then there will be no available theory-based reasons to weigh against his theory-neutral reasons. For such a person, theory-neutral reasons will be the *only* reasons applicable to how he ought subjectively to behave. However, if *strong* moral skepticism turned out to be demonstrably true—if it turned out that claims of the form "X is morally better than Y" were not the sort of claim that can be true, or that they were not the sort of claim whose truth we could ever hope to get information about—then theory-neutral reasons would, like the rest of morality, also be undermined. With no way to do moral philosophy, we would be left with little choice but to merely pursue our non-moral interests, with the thought that we would be subjectively no less likely to be acting morally by doing so than we would be by doing anything else. Conversely, if we found some sort of meta-ethical reason to believe that objective moral claims *do* have discoverable truth values, that would strengthen the reasonableness of following the recommendations of this dissertation even at the expense of non-moral interests. So the question of whether strong moral skepticism is true remains very important.

After strong skepticism, the next most important idea is that the best theory-neutral way to get at moral goodness is by using people's moral judgments as proxies. What if this turned out to be false? Suppose we somehow discover that people's *non-moral* desires track moral value better than their moral judgments do: that is, suppose that people find morally good states of affairs—even states of affairs which they have not consciously recognized as morally valuable—to be more pleasurable or more beautiful than the same people would find the same states of affairs to be if those states of affairs were not morally valuable. Picture Huck Finn, feeling a seemingly non-moral desire to help his friend Jim, despite mistakenly believing that helping Jim would be morally wrong.[74] Personally, I suspect that insofar as things like this happen in real life, they are examples of non-moral desires *happening* to align with true moral value, rather than of non-moral desires *tracking* value.[75] But one can tell stories in which genuine value-tracking is taking place. For example, perhaps behaving morally confers an evolutionary advantage, by decreasing our chance of offending potential allies. And perhaps, if unconscious moral reasoning or moral perception is possible, and if *conscious* moral judgments are more subject to rationalization, indoctrination, or other biases than subconscious ones are, it would be evolutionarily advantageous for some of our seemingly non-moral desires to be influenced by subconscious moral judgments while *not* influenced by conscious ones. If these admittedly-unlikely but not impossible ideas turn out to be true, then desires could turn out to be more value-tracking than moral beliefs. Anyhow, if we discovered that desires indeed *do* track moral value better than moral beliefs do, the Theory-Neutral Reason to Promote Recognition would vanish, the Theory-Neutral Reason to Promote Motivation would *invert*—we would find ourselves with a theory-neutral reason to encourage people to obey their allegedly non-moral desires at the expense of obeying their moral theories—and the Theory-Neutral Reason to Promote Success would change in subtle ways.

Aside from discoveries that completely disrupt my framework, there are many others which would alter the details. For example, if we accept that moral claims have discoverable truth values and that the best way to discover them is by attempting to engage in some sort of moral reasoning rather than by appealing to non-moral desires, the next question is *who* might be able to engage in such reasoning. I have been vague about this, using words like "moral agent" or "person". These terms showed up repeatedly in my discussion: only to the extent that a being is capable of moral discoveries do we have a theory-neutral reason to try to help that being recognize morally good consequences, feel motivated to act on that recognition, and succeed at that action, and only to that extent will the Neutral Policy treat that being's goals and rights as significant. It matters, therefore, who falls into this category, and to what extent. Various primates and cetaceans? Only adult humans? A tiny subset of adult humans? If we take theory-neutral reasons seriously, then trying to figure out exactly who is and is not a moral agent will be a high priority. Wasting resources on the fulfillment of non-value-tracking goals would be a mistake—except, of course, insofar as we have *theory-based* moral reasons to concern ourselves with such goals—as would neglecting the fulfillment of goals that *are* value-tracking. *How* moral truths can be discovered is also important, since one of the most fundamental theory-neutral reasons calls for us to facilitate such discovery. All of these are meta-ethical questions: to know who can discover moral truths, and how to help them do so, we need to figure out the fundamental nature of moral truth.

In addition to important meta-ethical questions about what moral truths are and how they can be discovered, the contents of our theory-neutral reasons could also be affected by new empirical information, especially in the realm of psychology and social sciences. We saw in Section 5.1 that the main theory-neutral reasons are stronger if people tend to be significantly motivated by their moral beliefs, and weaker if people mostly ignore their moral beliefs to follow their non-moral desires. Which of these scenarios hold is a question of empirical psychology, and worthy of our attention. Also, recall that we have a theory-neutral reason to try to increase people's degree of moral motivation. I speculated in Section 2.3.2 on how this might be done, but I am not a psychologist; the way it really ought to be done is via careful experimentation to find the best strategies for building character.

Much of this dissertation focused on our theory-neutral reason to help people, and establish policies that help people, achieve their goals. I suggested that we should try to maximize the supply of primary goods; that we should try to aim for a comfortable human population with enough people for division of labor but not so many as to make mere survival a full-time job; and that we should try to establish a liberal political system which permits individuals to pursue their own conceptions of moral goodness. All of these suggestions were grounded primarily on the claim that they would increase people's average degree of goal fulfillment. This is ultimately an empirical claim; it might turn out that average goal fulfillment is higher under alternative conditions. Perhaps hardship and overcrowding would, in the long run, lead to faster elimination of ineffective lifestyles. Perhaps a sufficiently benevolent dictatorship, making use of recent advances in surveillance and computation technology which were not available back when the Soviet experiment failed, could centrally manage the world's resources more effectively than they are managed under liberal capitalism. These possibilities seem remote, but cannot be ruled out analytically; only a study of history and of sociology can truly determine what kind of world is best at moral-preference-fulfillment.

On a smaller scale, there is plenty of room for further fine-tuning. Which types of goods tend to be most useful to arbitrarily-chosen human beings? What is the optimum population level for different regions? Which laws are best at protecting people's liberty and facilitating productive activities? All of these are empirical questions, not philosophical ones. Armchair philosophy cannot really tell us *how* to create a society in which a consequence's being morally valuable tends to make it more likely to come about; all it can tell us is *that* we should try to do so. I have sketched the theory-neutral reasons that are implied by my current sociological beliefs; if different sociological beliefs were justified, the particular recommendations I have suggested would be undermined. Theory-neutral reasons themselves would not be undermined: they would still be important, but would simply have a different content than I have claimed that they have.

In addition to the meta-ethical and empirical assumptions, I also had Section 1.2's more normative-looking claim that we ought to try to do what is subjectively right, where subjective rightness involves some sort of moral hedging strategy like the "expected rightness" calculus. What if we instead chose a different notion of subjective rightness? For example, suppose we replaced "perform whatever action has the highest expected rightness" with the more risk-averse "perform whatever action is least likely to be *seriously* wrong". The theory-neutral reasons I described above would still be applicable—faced with two options, either of which might be wrong in unanticipated ways, the fact that one empowers others to do good is still a reason to choose that one, since the indirect good it produces could help mitigate any wrongness it turns out to involve—but the precise way in which those theory-neutral reasons would be weighed against theory-based reasons would be different and probably weaker: it would be much easier for the fact that a single plausible moral theory deems an action to be terrible to outweigh the fact that many various moral theories judge that action to be slightly above-average. Meanwhile, under other possible strategies for decision-making under uncertainty, such as "perform whatever action is recommended by the objective moral theory you deem most likely"—which is the strategy one might adopt if one rejects entirely the idea of moral hedging, and insists that an action cannot be right unless the agent was specifically focused on the one true moral theory when choosing it—theory-neutral reasons would vanish entirely. So the choice of subjectivization strategy very much matters.

A final reminder is warranted. An agent, in addition to seeking new information pertinent to how to behave under moral uncertainty, should of course also seek new information pertinent to objective moral judgments—should try to reduce his moral uncertainty. The "should" here is moral: we have a theory-neutral reason to seek both kinds of information. A theory-neutral investigation of ethics absolutely *cannot* give the final word on right and wrong. All of the recommendations I have offered were parasitical on people's ability to make potentially-truth-tracking objective moral judgments and to act upon them, and this ability that would be lost if everyone exclusively followed theory-neutral reasons and ignored theory-based ones. So theory-neutral considerations must ultimately be one part of a greater moral system, combining with theory-based reasons. Fleshing out the *full* system would require examining candidate moral theories, formulating new ones not considered yet, and evaluating the strength of the theory-based reasons in favor of each. In the long run, perhaps we will have a science of ethics that is able to convincingly answer *all* moral questions, eliminating *all* moral uncertainty; on that day, theory-neutral considerations can be dropped from the system as obsolete. For now, however, it offers useful guidance about what we should do and how we should treat each other, while we await that happy day.

NOTES

[1] I owe the "lie" example, and the general point being made here, to Smith (2010), p. 91. Parfit (2009 draft) calls the "it is wrong to do what you believe to be wrong" view the "Thomist View" after Thomas Aquinas; see Section 21: "Acting in Ignorance".

[2] This case is meant to be structurally equivalent to the case given in Regan (1980), in footnote 1 of Chapter 11, which I first encountered as the "Mine Shafts" case in Parfit (1988 draft), p. 3. The only significant difference is that my version involves moral uncertainty rather than factual uncertainty. A more complex version which does involve moral uncertainty, and which illustrates the same basic point as mine does, appears in Lockhart (2000), p. 82.

[3] I do not mean to be quarreling here a Scanlonian buck-passing view—see Scanlon (1998), Chapter 2, Section 4—in which what really matters are the things upon which morality supervenes, not morality itself. If one wants one's actions to accord with the reasons upon which morality actually supervenes, *de dicto*, this will be sufficient to motivate my discussion; wanting them to accord with morality *per se* is not essential.

[4] See Smith (2010) for a discussion of how to subjectivize a moral theory.

[5] Lockhart (2000). By "those who have followed him" I have in mind primarily Jacob Ross and Andrew Sepielli, both of whom have influenced my discussion.

[6] See, e.g., Temkin (1987).

[7] Lockhart (2000) tries to solve some of these problems, but his solution is not entirely satisfactory. See Sepielli (2006) for what I take to be a decisive criticism of Lockhart's approach.

[8] Hume (1739), Book 3, Part I, Section I. Please note that I am *not* endorsing—indeed, I will be directly contradicting—Stephen Jay Gould's famous notion of "non-overlapping magisteria" in which factual premises have *no* relevance for moral questions—I am only acknowledging that a moral argument needs *at least one* moral premise *in addition to* whatever factual premises it might have.

[9] In particular I have in mind Rawls (1993) and the literature that has followed him.

[10] Jacob Ross makes essentially the same distinction when he suggests that under some circumstances we should believe one theory but accept for purposes of guidance a different theory. See Ross (2006), p. 743.

[11] Hume (1748), in Section X, famously answers "no" to this question: all purported miracles, including divine revelations, are better explained by appeal to the unreliability of the alleged witnesses.

[12] I have in mind Kant (1785).

[13] Derek Parfit makes an observation along these lines, and spells it out much more thoroughly, in Parfit (2009 draft), Section 40: "The Impossibility Formula".

[14] Rawls gives a clear description of this process, which he dubs "reflective equilibrium". Roughly: we form our theories, we compare their implications with our particular judgments, we respond to any disagreements by modifying the theories, modifying our particular judgments, or both, and then we repeat until there are no remaining areas of disagreement between the two. Rawls (1971), Section 4.

[15] For a more detailed discussion of possible sources of moral intuitions and of the implications of that for their reliability, see Singer (2005).

[16] As Kelly (2005) puts this point, learning others' views on a topic gives us "higher-order evidence" about the convincingness of the justification for our own views.

[17] Elga (2007) gives a plausible account of how much to adjust our confidence in the face of disagreement. He argues that my confidence in my judgment on a topic once I have formed that judgment and have learned other people's judgments on the same topic should be equal to my prior confidence that I would be right if we had that pattern of disagreement. If this is correct, then I cannot rationally maintain high confidence in my views, given disagreement about them, without also having an explanation for why it was much more likely that I would form the right view and the people disagreeing with me would form the wrong one than that I would form the wrong one and they would form the right one.

[18] I owe this argument to my father. It also echoes Steven Weinberg's famous quote, "With or without religion, good people can behave well and bad people can do evil; but for good people to do evil—that takes religion." I think he has misdiagnosed the problem slightly: any moral overconfidence, whether due to religious faith or some other poor moral epistemology, carries with it the seeds of disaster.

[19] Compare with: "[I]t is as certain that many opinions, now general, will be rejected by future ages, as it is that many, once general, are rejected by the present." Mill (1859), Chapter 2.

[20] For a philosophical treatment of split-brain studies and their implications for our own wholeness, see Nagel (1971). For more about the empirical evidence suggesting that even normal people have two minds, see Bogen (1986).

[21] See the classic argument in Singer (1972).

[22] For a more detailed account of possible disasters and of a few possible efforts at prevention, see Bostrom (2002).

[23] I take this to be consistent with the broad way many people use the terms "consequences" and "consequentialism". For example, Derek Parfit writes: "[W]e can still be, in a wider sense, Consequentialists. In this wider sense our ultimate moral aim is, not that outcomes be as good as possible, but that history go as well as possible." Parfit 1984, Section 10.

[24] For example, Kant (1785) writes at the start of Section 1 that having a "good will" is a prerequisite of being worthy to enjoy happiness or prosperity.

[25] Cohen (1997).

[26] A more detailed discussion of our pragmatic reason to assume that ethics is not a hopeless endeavor is given by Ross (2006); see especially Part 2. He is addressing moral nihilism rather than the kind of "believing that there are moral truths, but thinking we have no access to them" skepticism I am discussing here, but the same argument still applies.

[27] For an example of the kind of reasoning I have in mind here, see Korsgaard (1996) on appealing to the concept of autonomy.

[28] Again, see Rawls (1971), Section 4, on "Reflective Equilibrium".

[29] Such a view is defended by Hutcheson (1726), Treatise 2, Section 1.

[30] This analogy between moral intuitions and linguistic intuitions has been suggested before, e.g. in Rawls (1971), Section 9.

[31] Probably there is some underlying reason why this works, although J. L. Mackie applies it to morality more generally when he writes:

> Morality is not to be discovered but to be made: we have to decide what moral views to adopt, what moral stands to take. No doubt the conclusions we reach will reflect and reveal our sense of justice, our moral consciousness—that is, our moral consciousness as it is at the end of the discussion, not necessarily as it was at the beginning. But that is not the object of the exercise: the object is rather to decide what to do, what to support and what to condemn, what principles of conduct to accept and foster as guiding or controlling our own choices and perhaps those of other people as well.

Mackie (1977), Chapter 5, Section 1. Notice that he is not denying that moral truths exist or that they carry normative weight; he is just claiming that they are brought about by us forming beliefs about them.

[32] Compare with: "[W]hen I speak of the cognition or judgment that 'X ought to be done' [...] as a 'dictate' or 'precept' of reason to the persons to whom it relates, I imply that in rational beings as such this cognition gives an impulse or motive to action[.]" Sidgwick (1907), Book 1, Chapter 3, Section 3.

[33] For more on this distinction between these two ways in which moral considerations can enter one's psychological decision-making process, see Railton (1984), Section 6.

[34] For example, the founder of the modern Satanist movement writes: "[A]nything resulting in physical or mental gratification was defined as 'evil'[. ...] So, if 'evil' they have named us, 'evil' we are—and so what!" LaVey (1969), in the section on "How to Sell Your Soul". This appears to me to be indicating that when he says "do evil for evil's sake", he really means it as an abbreviation for "the things which have traditionally been labeled 'evil' are actually good, so do those things for goodness's sake".

[35] Sidgwick (1907) calls this "the fundamental paradox of hedonism". See Book 1, Chapter 4, Section 2.

[36] Mill writes:

> As it is useful that while mankind are imperfect there should be different opinions, so is it that there should be different experiments of living; that free scope should be given to varieties of character, short of injury to others; and that the worth of different modes of life should be proved practically, when any one thinks fit to try them. [...] Where, not the person's own character, but the traditions or customs of other people are the rule of conduct, there is wanting [...] the chief ingredient of individual and social progress.

Mill (1859), Chapter 3.

[37] Lloyd Thomas (1988) does not give this definition explicitly, but I am extracting it from Chapter 3: "Experimental Consequentialism", especially pp. 36-38. This is also where he claims that liberal rights are a precondition for learning what is valuable.

[38] Lloyd Thomas (1988) seems to acknowledge this point when he writes: "The reason why having more reasonable views of what is intrinsically good is good from anyone's point of view is that we would not wish to pursue illusory conceptions of what is intrinsically good." (From Chapter 3, in the section on "Is knowledge of what is good good?")

[39] For a more detailed discussion of this idea, see Buchanan (2002), pp. 142-144, who attributes it to Alexis de Tocqueville.

[40] It is a *very* old idea: Aristotle (350 B.C.E.) speculates (in Book 2) that virtue arises from habit—in which case it most certainly makes sense to *practice* resisting vice.

[41] Rawls (1971), Section 11.

[42] Other philosophers have also noted the fact that primary goods are not *always* useful to a person. For example, see Schwartz (1973).

[43] For a more detailed discussion of how additional information can be worse than useless, see Bostrom (2009 draft).

[44] Compare with the point made by Buchanan (1975) in his footnotes 15-16, in which he states that our reason for wanting primary goods need not be egoistic.

[45] I owe this case, and the general point, to Barry Loewer.

[46] See Sinnott-Armstrong (2006), Section 3, for an overview and citations.

[47] Dworkin (1978), Chapter 9, gives a classic discussion of how utilitarianism can be corrupted by "external" preferences.

[48] This case closely relates to one I encountered in a science fiction novel: Ringo (2008).

[49] A connection could be drawn here to the ancient doctrine of Stoicism, which also instructed people to ignore hedonistic desires in favor of reason.

[50] If the reader is wondering why I say "give weight to preferences in proportion to their likelihood of being motivated by recognition of the moral truth" rather than just "give weight to preferences in proportion to their likelihood of according with the moral truth", it is because the considerations which the latter captures and the former does not are all theory-based in the same sense that fulfilling goals with known contents was theory-based. I briefly discuss this issue in Section 2.4.3.

[51] This list is not meant to be exhaustive. For example, there may be ways to create moral agents other than by creating human children: with sufficient advances in information technology we might be able to manufacture intelligent computers which can function as moral agents; and with sufficient advances in bioengineering, or simply sufficient application of selective breeding and time, we might also be able to produce non-human animals possessing moral agency.

[52] A major source for my treatment here is Parfit (1984), Section 130: "Overpopulation".

[53] One might be tempted to render "a person's utility is positive if she is glad that *the universe* came into existence" as the more familiar "a person's utility is positive if she is glad that *she* came into existence". For hedonistic utilitarianism, the distinction is unimportant; if a person never exists, she will have no experiences of any kind, regardless of whether the rest of the universe exists. But the distinction matters for preference utilitarianism. In principle, if we allow non-self-interested preferences to count as morally significant, a person could have every single one of her preferences be fulfilled, and yet not have any opinion about whether it was good that she was born—all that is necessary is that her deeply-held preferences be for consequences which would have happened regardless. For example, suppose that a given subject's only morally-significant preference is that there be beauty in the world, and suppose that there would have been beauty in the world whether or not she had been there to see it. Surely a hard-core preference utilitarian would not want to ascribe zero utility to her in such a situation—she may not have a morally significant preference one way or the other about her own existence, but the morally significant preference she *does* have was fulfilled. So it seems to me that our test for what counts as positive had better appeal to the world's existence, not just to the subject's existence.

[54] From Mill (1861), Chapter 5, who attributes the quote to Jeremy Bentham.

[55] This connection has of course been drawn by others as well. See, for example, Riley (1990), who carefully argues that utilitarianism supports democratic institutions whenever information allowing interpersonal utility comparisons—e.g. "this action benefits Person A *more than* it harms Person B"—is unavailable or is too expensive to gather.

[56] For a discussion and proof of this point, see List and Goodin (2001).

[57] Plato (380 B.C.E.) introduces the idea in Book V.

[58] This *sort* of view is, in a different context, suggested by Hurka (1983).

[59] Possibly I should say "the problem" rather than "the problems": conflict resolution issues can be construed as issues about the distribution of unowned resources, with the unowned resource in question being "the power to decide how this conflict shall be resolved".

[60] Mill (1859).

[61] Mill writes:

> Who can compute what the world loses in the multitude of promising intellects combined with timid characters, who dare not follow out any bold, vigorous, independent train of thought, lest it should land them in something which would admit of being considered irreligious or immoral?

Mill (1859), Chapter 2.

[62] Mill again:

> [N]ot only the grounds of the opinion are forgotten in the absence of discussion, but too often the meaning of the opinion itself. [...] Instead of a vivid conception and a living belief there remain only a few phrases retained by rote; or, if any part, the shell and husk only of the meaning is retained, the finer essence being lost.

Mill (1859), Chapter 2.

[63] Compare with the argument in Mill (1859), Chapter 2, that the level of confidence we are justified in having in any belief which is protect by censorship cannot be high enough to justify that censorship. Mill is implicitly suggesting that one way in which we can be justified in increasing our confidence in a view is to see that it has survived attempted challenges.

[64] Mill (1859), Chapter 3.

[65] This narrowness of interest is most apparent in Lloyd Thomas (1988)'s statement that "it is choices between activities considered to be not wholly of instrumental value [i.e. considered to be of at least some intrinsic value] that are significant [to experimental consequentialism]." (From Chapter 4, p. 74.)

[66] Lloyd Thomas (1988), Chapter 3, section about "Command Liberalism", p. 50.

[67] Compare with: "[One] must be able to hear [counterarguments to the prevailing view] from persons who actually believe them; who defend them in earnest, and do their very utmost for them." Mill (1859), Chapter 2.

[68] Mill (1859), Chapter 1.

[69] See Parfit (1984), Chapter 16.

[70] Rice (1970).

[71] My source for this story is Gantz (1991), Chapter 4. William Marsh Rice died in 1900, but the racist clause he inserted into Rice University's charter lasted until 1964.

[72] That is to say, Locke (1689).

[73] P(Mx|Gx) = P(Mx|Jx)*P(Jx|Gx) + P(Mx|~Jx)*P(~Jx|Gx) = (P(Mx|~Jx)+b)*(P(Jx|~Gx)+a) + P(Mx|~Jx)*(1-(P(Jx|~Gx)+a)) = ab + b*P(Jx|~Gx) + P(Mx|~Jx). P(Mx|~Gx) = P(Mx|Jx)*P(Jx|~Gx) + P(Mx|~Jx)*P(~Jx|~Gx) = (P(Mx|~Jx)+b)*P(Jx|~Gx) + P(Mx|~Jx)*(1-P(Jx|~Gx)) = b*P(Jx|~Gx) + P(Mx|~Jx). P(Rx|Gx) = P(Rx|Mx)*P(Mx|Gx) + P(Rx|~Mx)*P(~Mx|Gx) = (P(Rx|~Mx)+c)*(ab+b*P(Jx|~Gx)+P(Mx|~Jx)) + P(Rx|~Mx)*(1-(b*P(Jx|~Gx)+P(Mx|~Jx))) = abc + bc*P(Jx|~Gx) + c*P(Mx|~Jx) + P(Rx|~Mx). P(Rx|~Gx) = P(Rx|Mx)*P(Mx|~Gx) + P(Rx|~Mx)*P(~Mx|~Gx) = (P(Rx|~Mx)+c)*(b*P(Jx|~Gx)+P(Mx|~Jx)) + P(Rx|~Mx)*(1-b*P(Jx|~Gx)+P(Mx|~Jx)) = bc*P(Jx|~Gx) + c*P(Mx|~Jx) + P(Rx|~Mx). So P(Rx&Gx)+P(~Rx&~Gx) = P(Gx)*P(Rx|Gx) + (1-P(Gx))*(1-P(Rx|~Gx)) = P(Gx)*(abc+bc*P(Jx|~Gx)+c*P(Mx|~Jx)+P(Rx|~Mx)) + (1-P(Gx))*(1-(bc*P(Jx|~Gx)+c*P(Mx|~Jx)+P(Rx|~Mx))) = (2*P(Gx)-1)*(bc*P(Jx|~Gx)+c*P(Mx|~Jx+P(Rx|~Mx)) + (1-P(Gx)) + abc, which is equal to 0.5+abc if P(Gx)=0.5.

[74] For this case and others like it, in which people's moral judgments were less value-tracking than their non-moral desires, see Bennett (1974).

[75] Arpaly and Schroeder (1999) discuss the Huck Finn case. Their reading of the example, unlike mine, is that Huck would not be helping Jim if it were not the right thing to do. However, they share my sense that there is an important division to be made between desires which track morality and desires which merely accidentally accord with it; they write that if Huck were moved by rebelliousness or "blind sympathy", he would not be praiseworthy.

BIBLIOGRAPHY

Aristotle. 350 B.C.E. *Nicomachean Ethics*.

Arpaly, Nomy and Schroeder, Timothy. 1999. "Praise, Blame, and the Whole Self." *Philosophical Studies*, Vol. 93, pp. 161-188.

Bennett, Jonathan. 1974. "The Conscience of Huckleberry Finn." *Philosophy*, Vol. 49, pp. 123-134.

Bogen, Joseph. 1986. "Mental Duality in the Anatomically Intact Cerebrum." *Bulletin of Clinical Neuroscience*, Vol. 51, pp. 3-29. Accessed at <http://www.its.caltech.edu/~jbogen/text/mental_duality.html> on April 30, 2008.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios." *Journal of Evolution and Technology*, Vol. 9, No. 1.

Bostrom, Nick. 2009 draft. "Information Hazards: A Typology of Potential Harms from Knowledge." Accessed at <http://www.nickbostrom.com/information-hazards.pdf> on March 27, 2010.

Buchanan, Allen. 1975. "Revisability and Rational Choice." *Canadian Journal of Philosophy*, Vol. 5, No. 3, pp. 395-408.

Buchanan, Allen. 2002. "Social Moral Epistemology." *Social Philosophy and Policy*, Vol. 19, No. 2, pp. 126-152.

Cohen, Joshua. 1997. "The Arc of the Moral Universe." *Philosophy and Public Affairs*, Vol. 26, No. 2, pp. 91-134.

Dworkin, Ronald. 1978. *Taking Rights Seriously*. Cambridge: Harvard University Press.

Elga, Adam. 2007. "Reflection and Disagreement." *Noûs*, Vol. 31, No. 3, pp. 478-502.

Gantz, Kerri Danielle. 1991. "On the basis of merit alone: Integration, tuition, Rice University, and the charter change trial, 1963-1966." Rice University. Accessed at <http://scholarship.rice.edu/handle/1911/13495> on May 31, 2010.

Hume, David. 1739. *A Treatise of Human Nature*.

Hume, David. 1748. *An Enquiry Concerning Human Understanding*.

Hurka, Thomas. 1983. "Value and Population Size." *Ethics*, Vol. 93, No. 3, pp. 496-507.

Hutcheson, Francis. 1726. *An Inquiry into the Original of Our Ideas of Beauty and Virtue*.

Kant, Immanuel. 1785. *Groundwork of the Metaphysics of Morals*.

Kelly, Thomas. 2005. "The Epistemic Significance of Disagreement." John Hawthorne and Tamar Gendler Szabo (eds.), *Oxford Studies in Epistemology*, Vol. 1, pp. 167-196. New York: Oxford University Press. Accessed at <http://www.princeton.edu/~tkelly/papers/disfinal.pdf> on November 30, 2007.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

LaVey, Anton Szandor. 1969. *The Satanic Bible*. New York: Avon.

List, Christian and Goodin, Robert E. 2001. "Epistemic Democracy: Generalizing the Condorcet Jury Theorem." *Journal of Political Philosophy*, Vol. 9, No. 3, pp. 277-306.

Lloyd Thomas, David. 1988. *In Defence of Liberalism*. Oxford: Basil Blackwell.

Locke, John. 1689. *Second Treatise of Government*.

Lockhart, Ted. 2000. *Moral Uncertainty and Its Consequences*. New York: Oxford University Press.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. Reading: Cox and Wymon.

Mill, John Stuart. 1859. *On Liberty*.

Mill, John Stuart. 1861. *Utilitarianism*.

Nagel, Thomas. "Brain Bisection and the Unity of Consciousness." *Synthese*, Vol. 22, No. 3/4, pp. 396-413.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, Derek. 1988 draft. "What We Together Do."

Parfit, Derek. 2009 draft. *On What Matters*. forthcoming from Oxford University Press.

Railton, Peter. 1984. "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs*, Vol. 13, No. 2, pp. 134-171.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.

Regan, Donald. 1980. *Utilitarianism and Co-operation*. Oxford: Clarendon Press.

Rice, Tim. 1970. Lyrics to "Superstar." "Superstar" was produced by Andrew Lloyd Webber and Tim Rice, and sung by Murray Head. Published a single in 1970 by Decca/MCA Records.

Riley, Jonathan. 1990. "Utilitarian Ethics and Democratic Government." *Ethics*, Vol. 100, No. 2, pp. 335-348.

Ringo, John. 2008. *The Last Centurion*. Riverdale, NY: Baen Books.

Ross, Jacob. 2006. "Rejecting Ethical Deflationism." *Ethics*, Vol. 116, No. 4, pp. 742-768.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge: Harvard University Press.

Schwartz, Adina. 1973. "Moral Neutrality and Primary Goods." *Ethics*, Vol. 83, No. 4, pp. 294-307.

Sepielli, Andrew. 2006. Review of Lockhart (2000), *supra*. *Ethics*, Vol. 116, No. 3, pp. 600-604.

Sidgwick, Henry. 1907. *The Methods of Ethics*, seventh edition.

Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy and Public Affairs*, Vol. 1, No. 3, pp. 229-243.

Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics*, Vol. 9, pp. 331-352.

Sinnott-Armstrong, Walter. 2006. "Consequentialism." *Stanford Encyclopedia of Philosophy*. Accessed at <http://plato.stanford.edu/entries/consequentialism/> on February 11, 2011.

Smith, Holly. 2010. "Subjective Rightness." *Social Philosophy and Policy*, Vol. 27, No. 2, pp. 64-110.

Temkin, Larry. 1987. "Intransitivity and the Mere Addition Paradox." *Philosophy and Public Affairs*, Vol. 16, No. 2, pp. 138-187.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI*. Singularity Institute for Artificial Intelligence. Accessed at <http://singinst.org/CFAI/> on January 24, 2008.