Normative Qualia and a Robust Moral Realism

By

Sharon Hewitt

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Philosophy

New York University

September, 2008

Thomas Nagel

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To an Irishman, a Scotsman, an Englishman, and a Breton

ACKNOWLEDGMENTS

I am very thankful to my advisors, Thomas Nagel, Sharon Street, and David Velleman, for their interest in this project, their generosity with their time and energy, and their always insightful criticisms. I owe a special debt of gratitude to Sharon, whose passion for metaethics first inspired my own.

Many other people have also contributed to the refinement of the ideas in this dissertation through discussion and written feedback, including Yuval Avnur, Ned Block, Dana Evan, Don Garrett, Béatrice Longuenesse, Colin Marshall, John Richardson, Jacob Rosen, William Ruddick, Karl Schafer, Jonathan Simon, and Peter Unger. Derek Parfit provided particular help on that portion of Chapter 5 that addresses his writings on the Repugnant Conclusion. Jeff Sebo has been an invaluable conversation partner, not least because he has pushed me to live out the commitments of my view in a more consistent way. And, of course, I am heavily indebted to my most devoted opponent: Melis Erdur.

I am grateful for the financial support I received at New York University through a MacCracken Fellowship and a Frankel Fellowship, and for the welcome and support I received as a *pensionnaire étrangère* at the Ecole normale supérieure.

I am thankful to Caleb Miller, whose ethics class first introduced me to the joys of philosophy and whose encouragement and able advising led me to graduate school.

Finally, I am deeply appreciative of the unfailing love and support of my parents, Herbert and Debra Hewitt, the encouragement and humor of my sisters, Rebecca Hewitt-Newson and Sarah Hewitt, and the ever-surprising companionship of Simon Biausse.

ABSTRACT

This dissertation formulates and defends a version of moral realism capable of answering the major metaphysical and epistemological questions other realist theories have not: namely, "What makes it the case that our concept of goodness objectively applies to certain things in the world?" and "How can we know to which things it objectively applies?"

To answer these questions, I propose a descriptive analysis of normative concepts: an analysis of intrinsic goodness and badness as phenomenal qualities of experience. I argue that all of our positive experiences share a common phenomenal quality that can only be accurately described in normative terms—as "goodness"—and that our negative experiences all share a phenomenal quality of badness. I claim that we acquire our concepts of intrinsic goodness and badness from our experience of these qualities, and that it is thus a conceptual truth that an experience that has one of these qualities is intrinsically good or bad.

I address Moore's Open Question Argument against such an analysis by arguing that, though the question of pleasure's goodness has an open feel, two things explain this: (1) the fact that we have a concept of all-things-considered goodness which depends not just on a thing's intrinsic goodness but also on its instrumental goodness, which is not knowable by reflection on the mere concepts involved, and (2) the fact that we easily mistake the goodness or badness we associate with pleasure for an intrinsic normative property of it.

I go on to explain how the *pro tanto* goodness and badness of phenomenal experiences justify judgment-independent claims about which states of the world as a whole ought to be promoted, all things considered. I argue that to pursue anything but the greatest total balance of good over bad phenomenal experience for all subjects would be arbitrarily to ignore the normativity of some of these experiences.

Finally, I defend the hedonistic utilitarian implications of this view against arguments that it conflicts with our moral intuitions, arguing that our intuitions are more consistent with the practice of hedonistic utilitarianism than is usually recognized.

TABLE OF CONTENTS

INTRODUCTION

The aim of this dissertation is to formulate and defend a version of moral realism that answers, at least in outline, the major metaphysical and epistemological questions that other recent realist theories have not: namely, “What makes it the case that our concept of goodness objectively applies to certain things in the world?” and “How can we know to which things it objectively applies?”

To motivate this project, I provide two preliminary chapters. Chapter 1 is addressed to those who see little interest in tackling the metaphysical and epistemological questions of realism because they believe that some version of antirealism can preserve all of our everyday moral values, discourse, and motivation, with more ontological simplicity. I argue that accepting or rejecting the metaphysical and epistemological commitments of realism actually has implications for one’s motivational structure: belief in realism reduces perspectival bias in one’s motivations in a way that belief in antirealism does not. This makes the case for antirealism less conclusive. If we value the reduction of perspectival bias, then we should be interested in seeing whether there might be a version of realism with plausible answers to the metaphysical and epistemological questions realism poses.

In Chapter 2, I explain why four types of currently defended realist theories—intuitionism, minimal realism, ideal-observer theory, and synthetic naturalism—are inadequate alternatives to antirealism. I follow this discussion with a list of four criteria that a realist theory must meet if it is to be a plausible alternative to antirealism. These are the criteria for what I call a metaphysically and epistemologically “robust” realism.

After these two preliminary chapters, I turn to formulating my realist view and defending it against metaethical objections. In Chapter 3, I explain the core claim of my realist view: that intrinsic goodness and badness just are phenomenal qualities of our experience and that our concepts of intrinsic goodness and badness actually come from our experience of these phenomenal qualities. I argue that all of our positive experiences (like pleasure and happiness) share a common phenomenal quality that can only be accurately described as “goodness,” and that all of our negative experiences (like pain and sadness) share a common phenomenal quality of badness. I go on to argue that, if intrinsic goodness and badness just are phenomenal qualities of experience, then the intrinsic goodness or badness of an experience that has one of these qualities is objective: i.e., it does not depend on our judgments about it.

Chapter 4 has two aims. The first is to defend the claim that there is a conceptual connection between positive and negative phenomenal qualities and normativity. I defend this view against G. E. Moore’s famous Open Question Argument against analytic descriptivism. The second aim is to show how the conceptual connection I posit allows my view to meet the four criteria for a robust realism.

In Chapter 5, I explain some further details of my view: namely, the way in which, using only facts about the intrinsic goodness and badness of individual phenomenal experiences, it can justify claims about which states of the world as a whole are best, all things considered. The conclusion is a classical utilitarian one, but I discuss questions the answers to which utilitarians often take for granted. I ask whether one's own normative phenomenology is truly normative for others and whether the value of normative phenomenology is strictly additive, and I answer both questions in the affirmative.

Having defended my view from a metaethical standpoint, I spend the final two chapters of the dissertation defending its substantive normative implications. While I find the considerations presented in the first five chapters decisive in recommending hedonistic utilitarianism as the correct normative ethical theory, many people believe that hedonistic utilitarianism is just too crazy to be true, and that, if one has to embrace it to be a realist, one ought to abandon realism. In order to continue defending my version of realism, therefore, in Chapter 6 I address concerns about the utilitarian aspect of the view, and in Chapter 7, I focus on objections to hedonism. In both chapters, I argue that, given certain very general facts about the situations we actually face, the requirements of hedonistic utilitarianism in the actual world do not diverge nearly as far from our common moral intuitions as is often thought. The version of realism I present actually provides a robust metaethical justification for many of our strongest moral convictions.

PART I

MOTIVATION FOR A ROBUST MORAL REALISM

CHAPTER 1

THE IMPORTANCE OF MORAL REALISM

Many antirealists have taken pains in the last few decades to argue that nothing terribly important is at stake in the debate over moral realism, implying that it's more of a technical issue of purely academic interest than a question of vital importance for all individuals who make moral judgments. I am thinking particularly of R. M. Hare's essay "Nothing Matters," of Chapter Six of Simon Blackburn's *Spreading the Word* (as well as of his "Errors and the Phenomenology of Value"), and of Allan Gibbard's *Thinking How to Live*.[1] In "Nothing Matters," Hare describes the "confusion" that has "led many people to suppose that there is some vital issue at stake between objectivists

[1] R. M. Hare, "Nothing Matters," in his *Applications of Moral Philosophy* (London: Macmillan, 1972), 32-47; Simon Blackburn, *Spreading the Word* (Oxford: Clarendon Press, 1984), Ch. 6, especially pp. 197-98; Blackburn, "Errors and the Phenomenology of Value," in his *Essays in quasi-realism* (Oxford: Oxford University Press, 1993), 149-65; Allan Gibbard, *Thinking How to Live* (Cambridge, Mass.: Harvard University Press, 2003), especially pp. 13-17, 267.

and subjectivists,"[2] and he calls the ethical realism debate "purely verbal."[3] According to Hare, "so-called 'subjectivists' and 'objectivists'…are saying the same thing in different words."[4] Blackburn continues the "verbal" theme, discussing moral quasi-realism within a book on philosophy of language. Blackburn writes,

> The problem is not with a subjective source for value in itself, but with people's inability to come to terms with it, and their consequent need for a picture in which values imprint themselves on a pure passive, receptive witness, who has no responsibility in the matter. To show that these fears have no intellectual justification means developing a concept of moral truth out of the materials to hand: seeing how, given attitudes, given constraints upon them, given a notion of improvement and of possible fault in any sensibility including our own, we can construct a notion of truth.[5]

Gibbard calls his book *Thinking How to Live* a "realization" of Blackburn's quasi-realist project. His goal is to show how our expressions of our plans about everyday questions of "what to do" actually turn out to have all of the major characteristics of realist moral discourse. The upshot of this project, according to Gibbard, is that no one need reject expressivism because it leads to moral skepticism and a rejection of all moral judgment. People will never be able to avoid making decisions about "what to do" that for all intents and purposes look just like moral judgments.

[2] Hare, "Nothings Matters," 45.
[3] Ibid., 41.
[4] Ibid., 40.
[5] Blackburn, *Spreading the Word*, 198.

If it’s true that belief in antirealism should have little to no effect on our ordinary moral judgments, this has the very positive consequence of allowing us to preserve the things we want from morality—our moral convictions as well as the possibility of ongoing moral reflection, conversation, and improvement—while allowing us to avoid the difficult metaphysical and epistemological problems that have always troubled moral realism. Embracing antirealism would seem to allow us to retain moral talk, moral attitudes, and moral social pressure, but do away with the philosophically embarrassing hypotheses of intuitional insight into a Platonic realm of moral truth, or moral particles which buzz around acts of torture and other instances of egregious evil. This no doubt forms a large part of antirealism’s appeal.

The goal of this chapter, however, is to cast some doubt on whether antirealism really can give us all of the advantages of realism without its metaphysical and epistemological disadvantages. The antirealists quoted above have gone a long way towards showing that antirealism has a way of interpreting our moral statements—even our statements about objectivity—such that making these statements does not commit us to crazy metaphysical and epistemological views. However, the fact that we *could* interpret these statements in this way does not mean that this is what most people *actually mean* when they make them. The fact that moral statements could be interpreted as mere expressions of attitudes or plans, for instance, does not mean that most people who make moral statements don’t actually intend to be asserting some sort of metaphysical claim. It seems quite plausible that many people who use moral language intend it—in however vague a way—to make claims about some objective moral standard.[6] It seems that many people use moral language with the feeling that their moral judgments are meant to reflect values that are actually *in* the acts and states of affairs that they pass judgment upon. When asked to elaborate on this feeling, the non-philosopher has usually been reduced to an appeal to God's authority or to the self-evidence of certain basic value judgments. Even philosophers have not been terribly successful at composing convincing answers to the question "What do you *mean* by saying your values are *objectively* right?" This has led Hare to make the following remark:

> [T]here is one thing that I can say without any hesitation at all—that I do not understand what is *meant* by 'the objectivity of values', and have not met anyone who does. I really think the terms 'objective' and 'subjective' have introduced nothing but confusion into moral philosophy; that they have never been given a clear meaning, and have frustrated all serious discussion of the subject.
>
> For suppose we ask, 'What is the difference between values being objective, and values not being objective?' Can anybody point to any difference? In order to see clearly that there is *no* difference, it is only necessary to consider statements of their position by so-called 'subjectivists' and 'objectivists' and observe that they are saying the same thing in different words.[7]

It's easy to sympathize with antirealists who conclude from realists' inability to express even the meaning of their position that there really is no substantial difference between what realists mean when they say "X is wrong" and what

[6] Antirealist J. L. Mackie makes a similar semantic claim in *Ethics: Inventing Right and Wrong* (London: Penguin, 1977).
[7] Hare, "Nothing Matters," 40.

antirealists mean by the same phrase. These antirealists, it may seem, have the virtue of honestly admitting that, without God or a realm of Platonic forms, there is no meaning added by saying normative judgments are “objective.” And yet I don’t think we should be so sure that we have exhausted all of the possible ways in which the thought of those who see their moral judgments as reflecting an independent standard could differ from the thought of someone who admits that there is no moral standard apart from one’s moral judgments themselves. It may be that a realist way of thinking differs from antirealist ways of thinking in such a way that it actually affects the nature of realists’ moral convictions and the future evolution of these convictions.

Of course, antirealists have already recognized one way in which realist thinking is different from antirealist thinking: the realist is led to seek out answers to metaphysical and epistemological problems that are of no concern to the antirealist. This difference is generally counted as a point in favor of antirealism. What I intend to show in this chapter, however, is that the metaphysical and epistemological commitments of realism are not just superfluous claims that should clearly be dispensed with in the interests of theoretical simplicity. I am going to argue that taking seriously the metaphysical and epistemological commitments of realism can actually have a positive effect on the evolution of one’s moral convictions, because they put normative as well as motivational pressure on one’s self-interested bias (as well as on other sorts of bias one may have) in a way that antirealist commitments do not. Thus antirealists cannot simply argue for their position by saying that it gives us all of the good things about realism without the metaphysical and epistemological baggage. That baggage does some normative and motivational work, and this gives us some reason to continue to look for a plausible way of defending the metaphysical and epistemological commitments of realism.

I. The realism/antirealism distinction

Before arguing that realism can give us something that antirealism can't, I should first be clear about how I understand the distinction between moral realism and antirealism. I label "realist" any metaethical theory that asserts that our normative judgments are made true or false by some normative fact independent of facts about our normative judgments themselves, and "antirealist" any metaethical theory that does not assert this. Normative judgments include not only explicit statements or beliefs about what is good or bad, or right or wrong, but also normative *attitudes* that are not necessarily explicitly formulated but may nevertheless be latent in our dispositions or demonstrated in our behavior. According to the realist, our normative judgments reflect normativity that exists "out there" in the world, while according to the antirealist, normativity only exists from the perspective of our judgments. According to the antirealist, there is nothing outside the entire set of our normative judgments which suffices to make them true or false.[8]

[8] Perhaps some readers will note that I have phrased my definition of realism so as not to distinguish between specifically moral normativity and other sorts of normativity—epistemic normativity, for

Now some antirealists do use the vocabulary of truth and falsity in referring to the status of their normative judgments. They can mean various things by saying a normative judgment is "true." For some antirealists, whom we might call "minimal expressivists," saying that a particular judgment is "true" does nothing more than express the speaker's agreement with the judgment, and saying that a judgment is "false" merely expresses the speaker's disagreement with it. There are many more sophisticated interpretations that antirealists give of truth and falsity, however.

Constructivists, for instance, say that the truth of a normative judgment consists in the fact that the judgment would be produced as the conclusion of a certain sort of rational procedure of thought, that it would, for instance, be a judgment we would make having reached a state of reflective equilibrium under conditions of full information. Among the crucial inputs to all such constructivist procedures, however, are some preliminary normative judgments or attitudes, what Sharon Street calls the

example. I have done this partly for convenience—it is less cumbersome to talk about things being "normative" than about them being "morally normative"—but also because I believe that all varieties of normative realism are ultimately connected.

I believe that the very same metaphysical and epistemological questions that plague moral realism plague any sort of realism about normativity, and that, in the end, the sort of judgment-independent reasons I describe and defend in this dissertation are the only judgment-independent reasons that exist. This means that any judgment-independent epistemic reasons, for example, are going to have to be derived from these reasons. Judgment-independent answers to questions about what one ought to believe are going to have to be arrived at in the same way as judgment-independent answers to other questions about what one ought to do.

Thus part of the reason that I have not adopted a more specific vocabulary is that I want to leave open to the reader the possibility of reading this dissertation as an exploration of the possibility of *any* variety of normative realism. But since I won't actually be defending here the connection between judgment-independent *moral* reasons and judgment-independent normative reasons in general, if the reader has reservations about this connection, I recommend that he or she simply understand 'normative' and 'normativity' as shorthand for 'morally normative' and 'moral normativity'.

"grounding set of normative judgments."[9] So, although in constructivism there is some complexity to the way in which normative judgments support the truth (or falsity) of other normative judgments, there remains the fact that there is no truth standard for normative judgments entirely independent of the set of all normative judgments. Thus constructivism is still a brand of antirealism.

Classifying views as realist and antirealist can get more complicated than this, however, as it does when we try to evaluate the status of quasi-realist views. Gibbard claims, for example, that the central thesis of realism—"Normative claims can be true or false, independent of our accepting them"—can be understood as an expression of a plan to act in certain ways even in those hypothetical situations where one will have a plan to act differently.[10] It's questionable whether one can coherently plan in such a way, but there is a fairly simple point that can be made about this view, without reviewing all of its intricacies. Quasi-realists have at their disposal many different techniques for interpreting 'truth', 'falsity', 'objectivity', and all of the other key realist terms in such a way as to give them meanings within the antirealist framework. What separates realism from all types of antirealism, including quasi-realism, is not the fact that it uses these terms, but the way in which realism actually sets up a truth standard for normative judgments that lies beyond mere judgments, attitudes, or plans.

[9] Sharon Street, "Constructivism About Reasons," in Russ Shafer-Landau, ed., *Oxford Studies in Metaphysics*, vol. 3 (New York: Oxford University Press, forthcoming).
[10] Gibbard, *Thinking How to Live*, 183, 186. See also his *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990), especially pp. 164-66.

Whenever antirealists employ the terms ‘true’, ‘false’, and ‘objectively’, they do so without ever giving these terms a reference that escapes the confines of our judgments.

Consider what Blackburn writes about the antirealist theory of projectivism. Projectivism posits that we project our attitudes onto the natural world and that this justifies our talking about goodness and badness as if they were objective properties. Blackburn writes,

> The utterance ‘whatever I or we or anyone else ever thought about it, there would still have been (causes, counterfactual truths, numbers, duties)’ can be endorsed even if we accept the projective picture, and work in terms of an explanation of the sayings which gives them a subjective source. The correct opinion about these things is not necessarily the one we happen to have, nor is our having an opinion or not the kind of thing which makes for correctness. The standards governing projection make it irrelevant, in the way that opinion is irrelevant to the wrongness of kicking dogs. The temptation to think otherwise arises only if a projective theory is mistaken for a reductionist one, giving the propositions involved a content, but one which makes them about us or our minds. They are not—they have a quite different role, and one which gives them no such truth-condition.[11]

Blackburn writes in the footnote to this section,

> [T]he metaphor of ‘projection’ needs a little care. Values are the children of our sentiments in the sense that the full explanation of what we do when we moralize cites only the natural properties of things and natural reactions to them. But they are not the children of our sentiments in the sense that were our sentiments to vanish, moral truths would alter as well. The *way* in which we gild or stain the world with the colours borrowed from internal sentiment gives our creation its own life, and its own dependence on facts. So we should not say or think that were our sentiments to alter or disappear, moral facts would do so

[11] Blackburn, *Spreading the Word*, 219. See also his *Ruling Passions* (Oxford: Oxford University Press, 1998).

> as well. This would be…endorsing the wrong kinds of sensibility, and it will be part of good moralizing not to do that.

Blackburn remarks that when we truly project our attitudes onto the natural world, these attitudes take on a life of their own. We view them as bonded to natural properties in such a way that we no longer see values as dependent on our psychological states. Once we have "projected" or objectified our values in this way, we will endorse such statements as "Were everyone to think kicking dogs was acceptable, it would still be wrong." Blackburn says that what we reject in endorsing such statements is the approach of someone who, when deciding how to act, simply consults the attitudes of others and behaves accordingly. According to the projectivist, talk about judgment-independent truth is just another expression of our sensibility. We have a negative attitude toward people who kick dogs, and in addition we have a negative attitude toward people who *would* kick dogs if everyone thought it was okay. On the projectivist picture, metaethical pronouncements are explained as expressions of attitudes toward behavior in counterfactual situations in which people hold different attitudes. Claims to objectivity are, according to Blackburn, "a proper, necessary expression of an attitude toward our own attitudes."[12]

Yet projectivism is still only quasi-realist, not fully realist, and this is because, though it uses all of the language of realism, it does so only after reinterpreting the metaphysical claims of realism as mere expressions of attitudes with no metaphysical

[12] Blackburn, "Errors and the Phenomenology of Value," 153.

implications. In defining antirealism, I said that an antirealist theory doesn't assert that there is anything outside the entire set of our normative judgments which suffices to make them true or false. A projectivist might say this definition doesn't include his view because he maintains that it's the pain caused to the dog that makes kicking dogs wrong, not the fact that anyone judges it to be wrong. But the fact remains that, on the projectivist picture, *without one's initial projection of a negative attitude towards pain onto the natural property of being in pain, the dog's being in pain would not provide any reason for believing the kicking to be wrong*. Now, granted, a quasi-realist isn't going to *say this*. A quasi-realist is going to *say*, "The pain caused by kicking a dog *is* a sufficient reason to believe that kicking dogs is wrong, a reason that is independent of anyone's judgments or projections of value." However, the fact that quasi-realists use the language of judgment-independence does not change the fact that their interpretations of judgment-independence claims are, in the end, judgment-dependent. Their interpretations of the foregoing statement mention only expressions of attitudes—for instance, of a plan to act in a certain way. They do not interpret it as actually making reference to a normative fact outside our judgments. A quasi-realist does not actually believe in judgment-independent facts the way a realist does. He just *sounds* like he does, by reinterpreting the meaning of realist-sounding statements so that they have no metaphysical import. Thus quasi-realists are still antirealists.

Before moving on to discussing the implications of the realism/antirealism distinction, let me quickly note one way in which I have purposely chosen *not* to draw this distinction. I have purposely chosen not to talk about "mind-independence," but rather about "judgment-independence." The term 'mind-independence' is currently popular in metaethics, but I believe it ought to be understood as a shorthand for 'judgment-independence'. That is, I believe 'judgment-independence' is a more precise term for what those who speak of "mind-independence" generally intend. I don't myself use the term 'mind-independence' because, given what I take to be the most plausible form of realism—a theory based on the intrinsic value and disvalue of certain forms of phenomenology—to define moral realism as a claim about the "mind-independence" of moral facts would be misleading.[13] On my view, moral facts are not mind-independent. They are quite dependent on whether people are in the mental states of pleasure or pain. And indeed, most mainstream realist views do take moral facts to be at least in part dependent on facts about pleasure and pain. However, on my view and on these other realist views, facts about the goodness of pleasure and the badness of pain (along with whatever other moral facts there may be) *are judgment-independent*.[14] These views say, for example, that there is a moral fact to the effect that pain is *pro tanto* bad, and this fact—this badness—exists independently of

[13] For a good example of a misleading characterization of realism, see Simon Kirchin, "Ethical Phenomenology and Metaethics," *Ethical Theory and Moral Practice* 6 (2003): 246. He defines mind-independent realism as claiming "that there are ethical properties in the world, properties that exist independently of our experience of them; their existence does not depend on any sort of human response (or belief, desire, interest, aim, etc.), nor do human responses influence in any way what kind of ethical properties, goodness for instance, belong to any particular situation." Kirchin doesn't mention that some human responses to the world—for instance, pain—do on most mainstream realist views influence the ethical properties of a situation.

[14] Russ Shafer-Landau makes the same point, using the term 'stance-independence' (which he says he owes to Ron Milo) to refer to what I call "judgment-independence." See his *Moral Realism: A Defence* (Oxford: Clarendon Press, 2003), 15.

whether anyone judges it to be there. I believe this belief in judgment-independence—not mind-independence—is the crucial difference between realism and antirealism, the difference that has potentially important consequences for the evolution of one's first-order normative judgments and the motivation one has to act on them.

II. Arguments that belief in antirealism shouldn't affect our first-order normative judgments

But why should belief in judgment-independence matter? Why should a belief that our normative judgments are made true by judgment-independent normative facts make a difference to our first-order normative judgments—to which things we judge to be good and bad, right and wrong—or to the motivation that we have to act on these judgments? To begin our investigation of this question, let's take a look at some antirealist arguments for the conclusion that there is *no* reason that moving from belief in realism to belief in antirealism should change one's values or behavior.

Antirealists often argue that, if one's values or behavior change as a result of coming to believe in antirealism, this is the result of a misunderstanding of the nature of antirealism. In "Nothing Matters," for instance, Hare describes the true case of a young high school graduate who, upon reading Camus' *The Stranger,* and its hero's pronouncement that "nothing matters," went into a deep depression: smoking, saying nothing during meals, eating little, and wandering for hours in the fields. Hare explains to us that this response was a result of mistakenly thinking that "mattering" was something that things did, rather than realizing that “the function of the word ‘matters’ is to express concern,”[15] concern that one can feel regardless of whether the hero of *The Stranger* did. Hare reports that when he explained this to the young man, he “ate a hearty breakfast the next morning,” apparently cured of his motivational ills.

Street also argues that antirealism should not diminish one’s level of moral conviction if properly understood. After presenting a sketch of her constructivist view, she writes,

> Upon hearing this, one might wonder: can I ever feel the same conviction about the value of family (or any other of my basic values that has a similar justification) if I am aware that its normativity stems merely from the contingent fact that I have this very strong unreflective tendency, combined with the fact that there’s no good reason to resist it and good reason to endorse and encourage it as judged from the standpoint of my other values? Isn’t this just admitting that it is not really *true* that I should value my family members?[16]

Street’s answer is a firm “no.” She replies that, on constructivism, one’s having a strong unreflective tendency to take something to be valuable, combined with support for this tendency from one’s other normative judgments, is just what it *is* for something to be valuable. In fact, she goes on to argue that no stronger sense of being valuable even makes sense. Something’s being valuable only makes sense from the perspective of some further values which act as a standard against which it can be judged. If one asks, “Why should I take anything at all to be valuable?”, one is

[15] Hare, “Nothing Matters,” 37.
[16] Street, “Evolution and the Nature of Reasons” (Ph.D. diss., Harvard University, 2003), 185.

> both posing a normative question and yet in the same breath stepping back from and suspending one's endorsement of all normative judgments, thereby robbing the question of the standards that could make the question make sense. One cannot step back from the *entire set* of one's interlocking normative judgments at once, and ask, from nowhere, whether this set is correct or incorrect. There are, and could be, no standards to fix an answer to this question.[17]

According to Street, questions of value can only be answered from the perspective of one's other values. But from this perspective, there are very definite answers as to what one ought and ought not to value, "such that you (at least if you're anything like most of us) couldn't *not* value your family members and *not* be making a mistake—not unless you suddenly became someone very different from who you are, someone you would barely recognize—someone with very different unreflective tendencies to value, and very different normative judgments of all kinds."[18] That is, the dispositions and the further normative judgments that you as a matter of fact have make it the case that you ought to value family, *according to your own standards*. Street continues, "The fact that the mistake is on your own terms, as determined by standards that are ultimately set by your own normative judgments, should in no way be undermining, and it will not be if one understands that the standards that determine the truth and falsity of normative judgments can only be set from within the standpoint of a valuing creature."[19] In other words, the fact that any mistake we could make in valuing would only be a mistake in relation to our other values should not lead us to

[17] Ibid., 201.
[18] Ibid., 185.
[19] Ibid.

see mistakes in valuing as any less serious. They are serious because they go against our very own standards, when these are the only standards that exist, or could exist. Thus we should not worry, according to Street, that it is not *really* true that we should value certain things. It is true "in the strongest sense that makes sense."[20]

Let me present one final argument that belief in antirealism should not change the substance of one's values, this one from Blackburn. In "Errors and the Phenomenology of Value," Blackburn acknowledges that "someone might suppose that only commitments that describe the constitution of the real world have any importance and that all others are better ignored: a projective explanation of morality may then diminish the attention that person is prepared to pay to it." He even adds that "[t]his latter attitude is quite common." But he goes on to say that

> it is not the [antirealist] explanation of the practice *per se* that has the sceptical consequence, it is the effect of the explanation on sensibilities that have been brought up to respect only particular kinds of thing. So when people fear that projectivism carries with it a loss of status to morality, their fear ought to be groundless, and will appear only if a defective sensibility leads them to respect the wrong things.[21]

That is, "such people have a defect elsewhere in their sensibilities—one that has taught them that things do not matter unless they matter to God, or throughout infinity, or to a world conceived apart from any particular set of concerns or desires, or whatever."[22] According to Blackburn, belief in antirealism does not itself cause people

[20] Ibid.

[21] Blackburn, "Errors and the Phenomenology of Value," 156.

[22] Ibid., 157.

to feel a reduced sense of moral conviction. They will have a reduced sense of conviction only if they have the wrong sorts of attitudes: attitudes to respect only things they believe to be “objectively” valuable.

The common thread running through the arguments of Hare, Street, and Blackburn is that loss of belief in things’ mattering independently of our caring about them doesn’t rationally prevent us from caring about them nonetheless. If there is no judgment-independent normative standard, then there is no judgment-independent normative standard compelling us only to care about things that are judgment-independently valuable. Thus there is no reason for us not to care about exactly those things that we cared about when we were realists.

This is quite right. Coming to believe in antirealism does not rationally force one to abandon any of one’s value commitments, and I am not going to argue that it does. What I am going to argue is that coming to believe in antirealism *removes* a rational requirement on our values that’s given by a belief in realism: the requirement that one’s values reflect what is judgment-independently valuable. While the disappearance of this requirement does not itself compel any change in one’s values, it does mean that there is no longer one previously strong counterweight to the influence of perspectival bias on one’s values, and on one’s motivations in general. (I take values to be a subset of motivations: they are motivations that one approves of—“whole-hearted desires,” in Harry Frankfurt’s terminology.) Without this counterweight, we can expect the influence of perspectival bias to be stronger on the motivations of an antirealist than on those of a realist, all else being equal.

I will break my argument for this conclusion into two sections. In the first, I will explain what perspectival bias is and why belief in realism rationally requires one to get rid of it but belief in antirealism does not. In the following section, I will explain how the existence of this rational requirement can actually cause the motivations of a realist to change in a way that they wouldn't if the same person were an antirealist.

III. Perspectival bias

It is relatively uncontroversial that human beings, as we actually are, are not equally motivated to attend to the interests of all other human beings. (Nor are we equally motivated to attend to the interests of all other creatures that have them.) This is not to say that most of us aren't motivated to attend to the interests of *some* human beings besides ourselves. It's also not to deny that some of us are motivated by the interests of *certain* others to at least the same degree that we are motivated by our own interests. And it's not to deny that some of us are motivated by the interests of *all* others to *some* degree. Nevertheless, our degree of motivation to attend to a person's interests seems to be affected to a non-negligible degree by certain facts about that person's situation with respect to us: e.g., by how close the person is to us physically and by how strong our emotional ties to them are. Our degree of motivation also seems to be somewhat affected by the temporal distance of the interests that could be promoted or harmed: we seem to be less motivated by far future interests than by present ones.[23]

The point is that our particular situation—our spatial, temporal, and emotional perspective—affects the way that the interests of different people at different times motivate us. The effect of our perspective on our motivations I call "perspectival bias." In asserting the existence of this bias, I mean simply to be making an empirical claim, hopefully one that will be uncontroversial.

I now want to consider what attitudes a realist and an antirealist ought to have to perspectival bias, given their metaethical views. A realist is going to have to acknowledge that differences in motivation produced by perspectival bias do not reflect differences in the objective[24] value of others' interests. This is because, according to realism, the value of something—e.g., the value of others' pleasure and pain—is independent of the person who is evaluating it or the perspective from which they are evaluating it. Now a realist might have reason not to want his motivations always exactly to reflect the objective value of everything. This might be too distracting: it might prevent him from promoting anything of objective value because he's always thinking of other valuable things he's *not* promoting. But to the extent that it does best promote objective value for one's motivations accurately to reflect this

[23] For an engaging discussion of these and other types of motivational bias, see David Hume, *A Treatise of Human Nature* (1740), Book 2, Part 3, Sections 6 and 7.

[24] Throughout the dissertation, I will often employ the term 'objective' as a synonym for 'judgment-independent', since the former is shorter and also lends itself to a more natural adverbial form.

value, a realist has to regard his perspectival bias as a bad thing. He has reason to reduce or eliminate it.

Now it might be suggested that one sort of realist would have no reason to reduce his perspectival bias. Consider a realist who believes in agent-relative reasons—i.e., believes that it is judgment-independently true that different people have reason to do or promote different things—and who takes his perspectival bias as reflecting facts about what *he* has an objective reason to do or promote. This seems to be a case in which a realist has no reason to reduce his perspectival bias.

This is not exactly true, however, because such a realist must consider what reason he has for thinking that his perspectival bias is reflecting judgment-independent, agent-relative reasons. He is going to have to try to reduce his perspectival bias in order to gauge, impartially, whether this bias has judgment-independent merits. And if the realist subsequently has reason to allow perspectival bias to return, it will not be because of his belief in *realism*, but because of specific evidence that one's being physically and temporally closer to a person, and more emotionally attached to them, are judgment-independent reasons for one to promote that person's interests more than someone else's. Belief in realism, on its own, gives one a reason to reduce one's perspectival bias, even though there could be other reasons not to reduce it.

Belief in antirealism, on the other hand, does not, on its own, give one a reason to reduce one's perspectival bias because, according to antirealism, there is no objective value in things that our perspectival bias could "distort." An antirealist is not aiming at his motivations' accurately reflecting any value existing independently of them, so if all of his motivations (i.e., all of his values, desires, attitudes, and other dispositions) are skewed in a particular direction, he has no reason to change this. Now that doesn't mean that antirealists couldn't nevertheless view certain sorts of perspectival bias as bad. It would be coherent for an antirealist to have an independent motivation to get rid of perspectival bias. But in the absence of belief in realism, the most natural reason to value getting rid of it—the fact that getting rid of it will make one better at promoting something whose value doesn't change with one's perspective—is gone.[25] And in the absence of this reason, if an antirealist doesn't just happen to be motivated to get rid of perspectival bias, he has no reason to do so.

I anticipate that there will be two sorts of objection to this claim: one coming from quasi-realists and the other from constructivists. I will deal with the quasi-realists' objection first.

A quasi-realist's view allows him to say, "Even if you don't happen to be motivated to get rid of perspectival bias, it's objectively bad, and you thus have a reason to get rid of it." But while this sounds like an objection to my claim, recall what the quasi-realist means by such statements. When a quasi-realist says, "You have a

[25] My point here parallels one that David Enoch develops at length in his article "Why Idealize?". Enoch argues that if one does not believe in response-independent normative facts, one can't appeal to the natural rationale for saying that it is only our "ideal" responses that are normative, not our actual, "biased" responses. The natural rationale, according to Enoch, is that only ideal responses reliably track the normative facts. See Enoch, "Why Idealize?" *Ethics* 115, no. 4 (July 2005): 759-87.

reason to A," or, "X is objectively bad," he is only expressing some very complicated attitude of his own toward A-ing or toward X: for instance, planning to A even in cases where he will not plan to A. The quasi-realist is not actually disagreeing with the *metaphysical* claim I made above: that there is no judgment-independent reason for him to change his motivations. Nowhere in his theory does the quasi-realist make a metaphysical claim affirming the existence of such a reason, rather than simply expressing his own attitudes towards certain motivations, and thus his position is entirely consistent with what I have said.

Let's turn now to constructivism. Recall that constructivists hold that one may have a reason to change one's values and/or other motivations if they do not stand up to a certain sort of rational reflection conducted in light of each other. A constructivist might propose that the proper sort of "rational reflection" is such that engaging in it actually leads one to give equal weight to all people's interests. A constructivist might even say that rationality itself demands that one give equal weight to all people's interests.

Whether we define rationality or rational reflection as requiring this or not, however, the crucial question is whether any particular individual has a reason to be "rational" in the sense defined. If the constructivist is to remain an antirealist, then she is not going to be able to posit any judgment-independent reasons. And thus, if someone doesn't care about being rational in her sense—if someone doesn't care about reflecting in a certain way, or if someone simply doesn't want to treat all people's interests equally, and the value of treating them all equally is not entailed by any other values the person holds—then an antirealist cannot assert that this person nevertheless has a reason to conform to this standard. If the constructivist builds anything else into the notion of rational reflection except the resolution of conflicts among the motivations one actually has (along with their strictly logical implications), she is departing from antirealism.

Now, as I noted before, it *could* be the case that someone does have independent motivation to get rid of perspectival bias. They might be naturally wired that way, or they might have acquired such a motivation through the influence of others. Perhaps it is even the case, as Christine Korsgaard seems to suggest, that it is a deep feature of human beings that they want to reduce perspectival bias, at least to some extent.[26] If one has this sort of independent motivation, then one will have a reason, even on antirealism, to try to strengthen one's motivation to look after those who are physically or emotionally distant from them. But constructivism itself does not give one such a reason. And realism gives one such a reason *in addition* to any natural motivation one might have.

[26] Christine M. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), 132-45, especially p. 145.

IV. Motivation to remove perspectival bias

In the last section, I asserted that our motivations are affected by perspectival bias, and I argued that belief in realism necessarily gives us a reason to try to get rid of this bias, one that belief in antirealism does not. However, this doesn't show that believing in realism will actually tend to reduce this bias more than believing in antirealism. Even if belief in realism gives us a reason to reduce it, this reason can't affect our motivations all by itself; we have to be motivated to act in accordance with it. Someone might argue in the following way: Either we are motivated to reduce perspectival bias or we are not. If we *are* motivated to reduce it, then we will do so even if we believe in antirealism. And if we *aren't* motivated to reduce it, why would believing in some "objective reason" to do so change that? The aim of this section is to show how belief in realism *could* motivate us to reduce our perspectival bias even when we wouldn't be motivated to do so if we were antirealists.

The problem is in seeing how a belief that we have a judgment-independent reason to eliminate our perspectival bias can motivate us if we don't already value eliminating this bias. The answer is that the motivation to eliminate our perspectival bias can be indirect. It can be derived from our motivations to promote other things—motivations that antirealists may share but that, without a belief in realism, will not be converted into motivation to eliminate perspectival bias.

If we believe that moral facts are independent of our perspective, then we will not believe that our motivations necessarily reflect the reasons we have to promote things. We will believe that our motivations could be misleading, and we will need some method for determining which are the reliable motivations and which are the misleading ones. One might think that a realist would just take his strongest motivations as indicating judgment-independent reasons and his weakest ones as more likely to be misleading, but someone who takes seriously the judgment-independence of reasons will not simply trust the strength of his motivations to tell him what these reasons are. He will consider what connection there is likely to be between judgment-independent reasons and his motivations. He will look at which of his motivations are likely biased by his own perspective or background, which of them are consistent over time, which of them cohere with each other, and which of them are consistent with the motivations of others. He will look for consistencies that indicate a pattern *beyond* the one that is produced by the peculiarities of his own viewpoint. It is the motivations that stand up to this kind of metaphysical and epistemological scrutiny that he is going to take as most likely to indicate judgment-independent reasons, and these will be motivations that have been corrected for perspectival bias.

This means that even a realist who has no independent motivation to get rid of perspectival bias is going to be making a distinction between motivations that are affected by perspectival bias and those that are not, because he takes only the latter to indicate judgment-independent reasons. We still have to show, however, that a realist will be more likely to be motivated by those motivations he takes to indicate judgment-independent reasons than by those he takes not to. Why would this be?

First of all, it could be the case that certain individuals just have a basic disposition to be more motivated by something if they think there's a judgment-independent reason for them to be motivated by it instead of merely a judgment-dependent reason. That is, certain people could be directly motivated by judgment-independence just the way that others are directly motivated by other properties, like sweetness, pleasantness, or beauty. In fact, all of us are motivated by judgment-independence in certain cases. Given the choice between reaching for a piece of fruit that we believe to exist independently of our perception of it, and a piece of fruit that we believe is only an optical illusion, we will choose the former (at least if we're hungry). And given the choice between pursuing a romantic relationship with someone whom we believe to have thoughts and experiences, and someone who appears in exactly the same way to us but whom we believe to be a zombie, most of us will again choose the former. So there are examples of cases in which we are motivated by a belief that something *is* a certain way more than by a belief that it merely *seems* to us that way. Something like this could explain why a realist who comes to believe that some of his motivations are veridical and some illusory will have more second-order motivation to respond to the former, when he wouldn't have this same motivation if he believed that all of his motivations were objectively on a par.

But perhaps it will be suggested that, while it makes sense to favor judgment-independent fruit over judgment-dependent fruit, and judgment-independent lovers over judgment-dependent lovers, there just doesn't seem anything to recommend judgment-independent *reasons* over their judgment-dependent counterparts. What if one just doesn't feel inclined to care about judgment-independence when it comes to reasons? Can a belief in realism be expected to have any effect on one's motivations in such a case?

It seems that it may, due to the way in which believing something changes the way that we represent the world to ourselves. If I believe, for example, that the walls of my friend's house are blue, then when I think of the walls of my friend's house, I will think of blue walls rather than walls of any other color. Perhaps if my friend has just repainted, I will slip up once or twice and think of the color that they previously were, until I remember her telling me that she repainted them. But when I remember her telling me this, I will try to picture the walls as blue, and over time my belief that they are blue will tend to make me think of blue walls when I think of her house.

Now consider a case involving a normative belief. If I come to the conclusion that all pleasure objectively has the value that it seems to have when I am experiencing it myself, then, when I think of the pleasure of others, I will start to call to mind what I believe to be a veridical perception of its value: my perception of my own pleasure's value. But if in this way I start to perceive the value of others' pleasure as I perceive my own, then the motivation that I have to look after my own pleasure will begin to be extended to the pleasure of others. I will become more and more motivated by the pleasure of others as I more and more consistently represent its value in the way I believe is accurate. Thus it’s not necessary that belief in a judgment-independent reason to care about others’ pleasure or to care about eliminating perspectival bias motivate me “from scratch.” Belief in judgment-independent moral facts leads to a transfer of the motivation one feels about certain cases to other cases that one believes to be objectively equivalent. And this “bleeding over” of the motivation one feels when one considers something from one perspective to the motivation one feels when one considers it from another results in one’s perspectival bias’ having less influence on one’s behavior.

What belief in judgment-independent moral facts is capable of doing, then, is taking someone who has no independent motivation to treat others’ interests equally with his own and causing him to perceive others’ interests in closer to the way that he perceives his own, simply because he has become convinced that their value is objectively similar, and this in turn causes him to come closer to being equally motivated by others’ interests and his own. Belief in the objectivity of reasons—and in the distorting effect of differences between his own perspective and that of others’—is the crucial factor in this change in motivation.

This doesn’t mean, of course, that any particular realist will have less perspectival motivational bias than any particular antirealist. There are other factors that influence one’s perspectival motivational bias: for instance, the extent to which one naturally puts oneself in the place of others, one’s skill at imagining situations very far or different from one’s own, and the extent to which those around one encourage the reduction of perspectival bias. Differences in these factors could cause any particular antirealist to be more impartial than any particular realist.

In addition, a distinction needs to be drawn between realists of the type I have depicted—realists who examine the evidence for the probability that different motivations of theirs reflect judgment-independent value—and what we might call "dogmatic" realists: realists who rely on the mere strength of their motivations or the authority of a third party to tell them what is judgment-independently valuable. Realism held to in this way—without taking seriously the metaphysical and epistemological questions it poses—is *not* any more likely than antirealism to lead one to a less biased set of motivations and could very well lead to one's being even more biased.

But if a realist and an antirealist are equal with respect to all other influences on their motivations, and are willing to consider the metaphysical and epistemological foundations of their views, the realist is nevertheless subject to an additional influence that reduces his perspectival motivational bias, due to the way in which he substitutes what he takes to be more veridical perceptions of value for less veridical ones. This is in addition to any direct motivation he may have to promote judgment-independent value over merely judgment-dependent value.

V. Conclusion

If belief in realism reduces perspectival motivational bias in ways that belief in antirealism does not, this gives us reason to reconsider the claim that antirealism can give us everything we want from realism. Whether we're realists or antirealists, if we value treating all people's interests equally—and it seems that all of us at least have reason to want others to treat *our* interests as equal to theirs—then this gives us reason to take it as a strike against antirealism that it is likely to exacerbate people's tendencies to give preference to those people who are nearer to them in space or time and to those with whom they have emotional ties.

This is not the only factor to be considered, of course. As antirealists have been correct to point out, the metaphysical and epistemological commitments of realism are a very important factor in determining whether or not we ought to endorse it. My aim in this chapter was merely to show that these concerns are not the only ones either. Realism and antirealism stand to have differing effects on our motivation, effects that are not simply due to a misunderstanding of antirealism, and if we are at all concerned about these possible effects, then we have reason to continue to investigate whether there may be a plausible way of answering the metaphysical and epistemological questions posed by realism. It is to these questions that I now turn.

CHAPTER 2

INADEQUATE REALISMS

In Chapter 3, I will sketch the contours of my own realist view, but first, I want to discuss several of the varieties of realism that have been defended by others. My goal in this chapter is to explain briefly why each of these versions of realism is unsatisfactory and in so doing to explain my motivation for developing the particular variety of realism that I do.

I believe each of the versions of realism I will discuss in this chapter has serious inadequacies, in metaphysics or epistemology or both. Many philosophers have concluded from the fact that there are consistently such problems in realist theories that problems of these types are endemic to realism. And yet while I agree that these inadequacies are fairly systematic, I don't think they are inescapable. Rather, I think that what is required is a radical shift in approach to theorizing about realism, one that takes very seriously the metaphysical and epistemological questions that realists have so long played down. Thus, in this chapter, I am not presenting other versions of realism because I hope to be able to patch them up or weave something out of materials drawn from them. Rather, I intend to make a fairly clean break with all of them and start off on what seems to me a radically different path. Nevertheless, it is the systematic problems in conventional approaches to realism that lead me to head off in the direction I do, and so it makes sense to examine, at least briefly, what is purposely being left behind.

In this chapter, I will examine four varieties of theory defended as realist: intuitionism, minimal realism, ideal-observer theory, and synthetic naturalism. I say these theories are "defended as" realist because, as will be discussed, there is reason to doubt in the end whether some of them are truly realist in the sense I defined in Chapter 1. Still, they are all of interest because their proponents take them to be plausible alternatives to antirealism.

By no means do these categories exhaust the field of existing realist theories, nor is it an infrequent occurrence that some realist theory falls into more than one of these categories or at least resembles more than one of them in certain respects. But despite the fact that this is not an exhaustive inventory of current realist views and that I can only offer here a fairly superficial discussion of each category, I believe this chapter will go some ways towards explaining where the primary deficiencies of current realist theories lie. My discussion of these theories will lead, in the final section of the chapter, to a list of four criteria which it is essential for a realist theory to meet if it is to succeed where these others fail.

I. Intuitionism

I will begin by discussing the most classic and most criticized of realist views: intuitionism. The philosopher best-known for having been a moral intuitionist is G. E. Moore,[27] but other self-described intuitionists have included Henry Sidgwick,[28] C. D. Broad,[29] W. D. Ross,[30] and A. C. Ewing.[31] Intuitionism, as a realist metaethical theory, posits that at least some moral facts are known by way of a special faculty of moral intuition. Now, while the theory is named for this epistemological stance, its epistemology implies a metaphysical position as well. Since intuitionism posits that one comes to know moral facts through a unique faculty of intuition, it implies that these moral facts are not the sorts of things with which one could become acquainted through any of one's other faculties: for instance, through the five senses or through one's faculty of reason (the "faculty of reason" being that which enables us to understand logic, not a faculty that tells us what ultimate—not simply instrumental—reasons we have for action). This implies that moral facts are not natural facts, nor are

[27] G. E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1903).
[28] Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Chicago: University of Chicago Press, 1907).
[29] C. D. Broad, *Five Types of Ethical Theory* (London: Routledge & Kegan Paul, 1930), 112, 264-73.
[30] W. D. Ross, *Foundations of Ethics* (Oxford: Clarendon Press, 1939), Ch. 8.
[31] A. C. Ewing, *The Definition of Good* (London: Routledge & Kegan Paul, 1948); and *Ethics* (London: The Macmillan Company, 1953).

they the deliverances of pure logic. According to intuitionism, moral facts are a metaphysical type unto themselves, and for this reason, require a unique epistemology.

The uniqueness of moral facts presents some problems. It makes it difficult for us to develop any very enlightening description of the nature of moral facts, since we cannot relate them to anything else on which we have a tighter conceptual grip. It also makes it difficult for us to explain the nature of the faculty of moral intuition. Unlike the scientists who continue to offer increasingly detailed accounts of the functioning of our physical senses, our computational abilities, and our memory, intuitionists seem unable to produce even the most basic description of the functioning of moral intuition. But then again, how could they be expected to produce such a description if the faculty is of a non-empirical nature?

The proposed uniqueness of moral intuition does not make its existence impossible, of course. Usually we do prefer to have theories which unify our explanations of various phenomena rather than separate them, and the fact that intuitionism makes moral knowledge stand alone is certainly a strike against it as far as explanatory utility and aesthetic value are concerned, but for all that, moral intuition might still exist, as odd as it would be. What seems a more definitive strike against intuitionism (though it still doesn’t amount to making moral intuition an impossibility) are its difficulties in providing an explanation for the fact that people are often in great disagreement with one another on moral subjects (as well as in disagreement with themselves over time), an explanation that does not undermine the ability of intuition to justify realist moral belief. On the assumption that there is a truth about moral matters, the existence of disagreement means that some of us are getting things wrong. But if we are supposed to have a special faculty of moral intuition that is a reliable guide to the nature of judgment-independent moral facts, how have so many of us ended up going astray, and how do we know which of us these are? If intuitionism is to justify our belief that our moral judgments reflect a judgment-independent fact of the matter, it is going to have to provide some tools for distinguishing the veridical deliverances of a reliable faculty of moral intuition from beliefs which merely *seem* to be the products of such a faculty.

Intuitionists have frequently appealed to a feeling of self-evidence to justify belief in the objective truth of their moral judgments. The primary problem with this route is that very often both parties in a moral disagreement feel that their positions are self-evident. In such cases, feelings of self-evidence can't be used to distinguish the true beliefs from the false ones. Appeals to self-evidence can also be dangerous given the facility with which human minds are able to endow with a feeling of self-evidence such things as racial and social prejudices. Feelings of self-evidence seem more closely correlated with a failure or refusal to see things from other points of view than with a superempirical faculty of insight. As Peter Railton notes, "It is too easy for us to give a non-justifying psychological explanation of the existence in certain English gentlemen of something which they identified upon introspection as a faculty of moral insight, an explanation that ties this purported faculty more closely to the rigidity of prevailing social conventions than to anything that looks as if it could be a source of universal truth."[32] Unless intuitionism offers some details about the functioning (and what must often be the malfunctioning) of the special faculty it postulates, it is hard to see how it could be taken to be a serious alternative to antirealism.

Indeed, few philosophers explicitly defend an intuitionist view at present, probably due to these salient problems. At the same time, some realists do appeal to something very much *like* intuitionism. For example, Russ Shafer-Landau, in *Moral Realism: A Defence,* appeals to the existence of some self-evident moral principles in an effort to combat skepticism about moral knowledge and thus bolster the overall case for moral realism. Shafer-Landau defines a proposition p as "self-evident" if it is such that "adequately understanding and attentively considering just p is sufficient to justify believing that p."[33] He argues that we are justified in believing in some of our most basic moral principles in just this way, i.e., without any corroborating evidence or inference from other principles. He believes this is important because, if the moral realist has this sort of justification, then he can escape skeptical worries about circular reasoning and infinite regresses.

Let's consider how Shafer-Landau argues for the existence of self-justifying propositions. (I use the term "self-justifying" instead of Shafer-Landau's "self-evident" because it seems to reflect better the property he describes and will make the

[32] Peter Railton, "Moral Realism," *Philosophical Review* 95 (April 1986): 163-207, p. 206.
[33] Russ Shafer-Landau, *Moral Realism: A Defence* (Oxford: Clarendon Press, 2003), 247.

following discussion clearer.) By Shafer-Landau's own admission, "[t]he best that can be done here is to offer candidates that are appealing (if any are), and to reply to criticisms of the idea."[34] I agree that it's difficult to think of any other way in which the self-justifying nature of certain propositions could be defended, since what is being rejected is the need to resort to independent evidence or reasoning to support their justification. Indeed, the best, and perhaps the only, evidence that a proposition is self-justifying is that it strongly and reliably strikes us as being so. In this spirit, Shafer-Landau offers the candidate propositions that, "other things equal, it is wrong to take pleasure in another's pain, to taunt and threaten the vulnerable, to prosecute and punish those known to be innocent, and to sell another's secrets solely for personal gain."[35] Unfortunately, as compelling as the truth of these principles may seem to most of us, not everyone believes them to be true. This means that one can raise to Shafer-Landau the objection from disagreement that I already raised to intuitionist views in general. If we allow that there are people who do not find these principles strongly compelling, even after having adequately understood and attentively considered them, does this not undermine our justification (1) for believing that these principles are self-justifying and/or (2) for believing in the principles themselves? Shafer-Landau addresses these two worries in turn.

[34] Ibid., 247.
[35] Ibid., 248.

The first worry is that, if there are some people who, though they have adequately understood and attentively considered a proposition, do not believe it, then it cannot be the case that adequately understanding and attentively considering the proposition is sufficient to justify believing it. The implied premise is roughly that, if a proposition were self-justifying, then everyone would believe it. Shafer-Landau rejects this premise. He argues that one may fail to assent to a self-justifying proposition for a number of reasons. One might have some very extreme sort of psychological malfunction, "a breakdown in [one's] belief-forming mechanisms," that could go so far as to prevent one even from assenting to so-called analytic statements that one fully understands. Shafer-Landau argues that, "[e]ven for those propositions we take to be analytic (if any are), there is no logical or metaphysical necessity linking an agent's understanding of them and her belief in them."[36] However, we probably can't assume that the belief-forming mechanisms of most of the people who disagree with us about the status of basic moral principles are as far gone as this. Thus, for most cases of disagreement, Shafer-Landau offers a more mild diagnosis: "[A]n agent's other emotions or beliefs may stand in the way of accepting or believing a self-evident proposition. … Such impediments have various sources—gullibility, lack of experience, brainwashing, morally impoverished upbringings, facile thinking, etc."[37] That is, although certain moral propositions compel the assent of most reflective moral

[36] Ibid., 262.
[37] Ibid., 262.

agents, they may not compel the assent of all such agents due to distorting influences on some agents' belief formation.

Although Shafer-Landau does not himself appeal to the following analogy, I think we may find some support for his conclusion by considering how distorting influences could prevent assent to self-justifying propositions in the case of everyday visual perception. Epistemologists often grant that having a visual experience as of a bird flying by, for example, is itself enough to justify believing that a bird is flying by, at least in the absence of clear defeaters to this belief. That is, visual perception offers justification for belief entirely on its own merits, without appealing to any additional evidence or reasoning. As clear as it seems that we are justified in believing our perceptions purely on their own merits, however, it is certainly possible that one might *not* believe a particular perception. This might be because one believes one has a defeater for such a belief, such as the fact that one has just taken a hallucinogenic drug. On the other hand, one's failure to believe might be for some less epistemically respectable reason such as that one is part of a cult that teaches that there are no birds. Or one may wish so desperately that there be no birds that one can't bring oneself to believe that one has just seen one. In any case, I agree with Shafer-Landau that the mere fact that some people don't believe a certain proposition does not mean that they are not in possession of self-justifying evidence for that proposition. A proposition may, all on its own, be sufficient to justify belief in it, but someone may not believe it due to other influences on their belief formation. Disagreement over a proposition thus does not directly prove that it is not self-justifying.

But this is not the only worry provoked by the existence of disagreement over basic moral principles. The second worry that Shafer-Landau discusses is that disagreement might undermine, not our belief that some moral propositions are self-justifying, but our own justification for believing in these propositions. Intractable disagreement may serve as a defeater of justification we would otherwise have.

I want to point out, first, that Shafer-Landau's formulation of self-justification does not actually allow for this possibility. Because Shafer-Landau says that a self-justifying (in his terminology, "self-evident") proposition is such that adequately understanding and attentively considering it alone is sufficient to justify belief in it, it is not the case that disagreement could take away our justification for belief in a proposition while leaving its status as "self-justifying" untouched. That is, because Shafer-Landau's definition says adequate understanding and attentive consideration are *sufficient* for justification of belief in a self-justifying proposition, *there is no room for defeaters of belief in a self-justifying proposition*. If there are compelling arguments that we are not justified in believing certain propositions, then these are also compelling arguments that these propositions are not self-justifying according to his definition. This fact is obscured by Shafer-Landau's use throughout his discussion of the term "self-evident" rather than "self-justifying"; one may forget that self-evidence has been defined in terms of justification, and in particular, sufficient justification. What this means is that Shafer-Landau's replies to this second worry are even more important than he realizes. If he cannot justify our belief in certain basic moral principles in spite of disagreement, then he also has to give up belief that these moral principles are self-evident, according to the definition he has given. And if he can't offer a plausible *revised* definition, one which allows for potential defeaters of "self-justifying" propositions while preserving their justificatory force, then he has lost his hoped-for defense against moral skepticism.

Let's examine Shafer-Landau's reply to the worry that disagreement serves as a defeater of one's justification for believing certain basic moral truths. (We'll return to the question of revising Shafer-Landau's definition of self-justification.) Shafer-Landau's first response is a concession that, under certain circumstances, the presence of disagreement may very well serve as a defeater.[38] He elaborates somewhat later,

> It is true that awareness of disagreement regarding one's moral endorsements may serve as a defeater. It will do so if one has nothing to say on behalf of one's moral views, after receiving or conceiving of a challenge from a dissenter whose conflicting views are themselves coherent, compatible with the non-moral evidence, etc. Crucially, one is in a different epistemic position before and after confronting such disagreement. Prior to this sort of confrontation, one may be justified in one's belief simply because of having understood a self-evident proposition. But after the challenge is issued, one is required to defend oneself.[39]

That is, Shafer-Landau is admitting that a certain kind of disagreement about a proposition, though it does not necessarily mean one cannot be justified in one's belief

[38] Ibid., 263.

[39] Ibid., 265.

in that proposition, *does* mean that one cannot be justified *solely on the basis of adequately understanding and attentively considering the proposition*. He admits that the feeling of self-evidence we may have about a proposition is, in light of disagreement, insufficient to justify belief.

This is a larger concession than Shafer-Landau seems to realize. What it means is that, in any world in which disagreement about basic moral principles does exist, Shafer-Landau's appeal to "self-evidence" is doing *no justificatory work*. He has actually conceded to the argument from disagreement, conceded that in light of disagreement, one is required to defend one's beliefs by traditional methods: e.g., by citing other judgments which corroborate them and by identifying the sources of others' error. Shafer-Landau spends the rest of the section arguing that these more traditional sorts of support can provide justification even in the light of intractable disagreement.

Let's take a quick look at these arguments. Shafer-Landau's goal is to persuade us that one may be justified in one's moral beliefs if one has evidence and reasoning to support these beliefs, even if one is unable to convince others of their truth. He attempts to do this by offering three examples of non-moral cases in which he believes us to be justified in our beliefs despite our inability to convince others. The first case is that of one's justification in believing that the earth is round despite the fact that all of the evidence one may gather will not convince members of the Flat Earth Society. The second case appeals to one's justification in believing what one perceives or remembers perceiving, despite one's inability to produce evidence that convinces others of what one perceives or remembers. Thus far, I am in agreement with Shafer-Landau. It seems we are justified in believing that the earth is round despite the existence of the Flat Earth Society, probably because there is a good story to tell about how these people rely on faulty belief-forming methods; particularly, they don't seem to correctly employ the principle of inference to the best explanation. As for the cases of perception and memory, there it seems we have an obvious explanation of why our beliefs differ from others' and why we cannot bring them to agree with us: we are in possession of different bodies of evidence, because we have experienced something that they have not, and there is no way to make up for this difference.

It is not clear that these cases are sufficiently similar to the case of moral belief, however, since in these cases we can point to either an empirical reason why our bodies of evidence differ or a specific principle of reasoning that those who disagree with us have ignored. In the moral case, there seems to be no such straightforward explanation for our disagreement. Shafer-Landau's final example is more similar to the case of moral belief than the first two, but it is here that I find his claim to justification weakest. Shafer-Landau asks us to consider the position of a philosopher working on the problem of free will.

> Take someone who spends a career trying to solve the free will problem. She crafts a superb book (acknowledged as such even by her detractors), knowing all the while that she has nothing like a non-controversial demonstration or proof of her major claims. She is able, to her satisfaction, to respond to objections, to corroborate her

> conclusions with much supporting argument and evidence, and to offer diagnoses of her opponents' vulnerabilities. Yet all of this will fail to convince many or most of her colleagues. And she knows that. But I don't see that this reception forces her to suspend judgement on the matter she has thought so carefully about. Of course, others who have worked as hard and as well on the subject, but who disagree with her, are also justified in their views, and not all of these views can be true. But justification does not entail truth. What we have here, as elsewhere, is a case in which one cannot convince rational, open-minded, well-informed agents of the truth of one's own views. But, as we have seen, such views may be true nonetheless, and (the important point for present purposes) we may be justified in thinking them so.[40]

I am not at all sure that such a person is justified in her views. Perhaps she is not entirely unjustified, but surely she does not have the level of justificatory support for her beliefs that someone does who *saw* something happen and is simply unable to convince others of this fact. Our philosopher should surely be given some pause by the fact that she is acquainted with other philosophers whom she believes to be just as intelligent and dedicated to their research as she is and who nevertheless hold opposing views. Why is it that these others are not convinced by her arguments, nor by her diagnoses of their errors? If there is no difference in their abilities, nor in the evidence available to each of them, should this not make her wonder, even a little bit, whether she should refrain from believing anything about the issue, or even whether the issue itself might be misconceived? We can certainly grant Shafer-Landau that "justification does not entail truth": that a view may be true in spite of one's inability to convince others of it. The question is rather whether any individual, faced with the

[40] Ibid., 264.

persistent disagreement of those she believes to be in equally good epistemic positions, is justified in believing that she’s the one who’s gotten things right in this case, or even that there is a truth on this particular matter.

In the next section, on minimal realism, I will further discuss the status of disputed philosophical truths. For now, it is enough that Shafer-Landau recognizes that, in the wake of disagreement, our moral beliefs are in need of some defense. Whether this defense is successful or not is going to depend on the merits of the corroborating evidence we can compile in a particular case. The difficulty in ethics has always been finding anything that seems to count as corroborating evidence for our fundamental moral beliefs, and finding plausible, non-undermining explanations as to why others refuse to accept this evidence. But whether this search for a justification for moral belief in spite of disagreement will be successful or not (and I actually believe that it will), the point to be noted is that appeals to self-evidence or self-justification are *not* sufficient to justify our belief in moral propositions in the presence of disagreement.

Let’s return, however, to Shafer-Landau’s proposal that, even if we might not be justified in believing in self-justifying moral propositions because of the defeater of moral disagreement, they might for all that still be self-justifying and thus provide a defense against circularity and regress arguments for moral skepticism. If these two states of affairs are to be compatible, though, remember that it’s going to be necessary to modify the definition of self-justifying propositions which Shafer-Landau gave, in order to allow for defeaters. His original definition was this: "A proposition p is self-evident $=_{\text{df.}}$ p is such that adequately understanding and attentively considering just p is sufficient to justify believing that p." Let's add a clause allowing for a defeater arising from disagreement: "A proposition p is self-evident $=_{\text{df.}}$ p is such that adequately understanding and attentively considering just p is, in the absence of disagreement, sufficient to justify believing that p." Does the existence of propositions that fit this definition provide a defense against moral skepticism? Not in the presence of disagreement. *In the presence of disagreement, even a self-evident proposition does not give one sufficient justification for believing it.*

We thus see that, if intuitionism is going to provide a plausible alternative to antirealism in the presence of moral disagreement, then it is going to have to forget appeals to self-evidence and go into some detail about just how the special faculty of moral intuition functions and why it malfunctions in certain cases. Thomas Nagel, though he does not claim for himself the label "intuitionist," does offer some suggestions about moral epistemology which could bolster the intuitionist's case. He writes,

> Even where there is truth, it is not always easy to discover. Other areas of knowledge [besides ethics] are taught by social pressure, many truths as well as falsehoods are believed without rational grounds, and there is wide disagreement about scientific and social facts, especially where strong interests are involved which will be affected by different answers to a disputed question. This last factor is present throughout ethics to a uniquely high degree: it is an area in which one would expect extreme variation of belief and radical disagreement however objectively real the subject actually was. For comparably motivated

> disagreements about matters of fact, one has to go to the heliocentric theory, the theory of evolution, the Dreyfus case, the Hiss case, and the genetic contribution to racial differences in I.Q.[41]

The idea is that our personal interests and desires may offer an explanation for our disagreements about moral questions in spite of some sort of access we all have to the truth about the matter. Our own interests have been known to cause disagreement about matters of empirical fact, so why should we be surprised if they cause disagreement on a subject that is even more closely related to them and thus even more potentially threatening to them: truths about what we ought to do? The distorting influence of personal interests might be thought to offer an excellent way for the intuitionist to explain moral disagreement even in the presence of a special faculty of moral intuition.

But while the availability of such an explanation does certainly help to make intuitionism a bit more plausible, one might still wonder whether, given all of the actual moral disagreement that exists, there is any positive reason to believe that, underneath all of our clashing personal interests, there is a core of intuited moral truth. Nagel suggests that,

> [a]lthough the methods of ethical reasoning are rather primitive, the degree to which agreement can be achieved and social prejudices transcended in the face of strong pressures suggests that something real is being investigated, and that part of the explanation of the appearances, both at simple and at complex levels, is that we perceive, often inaccurately, that there are certain reasons for action, and go on to

[41] Thomas Nagel, *The View from Nowhere* (New York: Oxford University Press, 1986), 148.

> infer, often erroneously, the general form of the principles that best account for those reasons.[42]

Nagel is suggesting that the fact that moral agreement can be reached by parties who previously disagreed is an indication that both of those parties do have access to a truth of the matter, even if it often takes a great deal of work to break this truth free from the interference of their personal interests. It seems there are ways of "seeing past" the distorting influences on the formation of our moral judgments, most notably by comparing our judgments with one another's and especially with the judgments of persons whom we consider to be most impartial and best at moral reasoning.

John Rawls, for instance, proposes just such a procedure for discovering "reasonable principles" for deciding moral questions.[43] He outlines the characteristics of the class of competent moral judges: intelligence, knowledge of relevant non-normative facts, reasonableness, open-mindedness, awareness of one's own tendencies to bias, and ability to imagine oneself in the place of others. And he tells us which judgments of such judges are most reliable—those made in cases which satisfy all of the following conditions:

(i) the judge has no personal interest at stake

(ii) the case is actual, not hypothetical

(iii) the case has been carefully investigated by the judge

[42] Ibid.

[43] John Rawls, "Outline of a Decision Procedure for Ethics," *The Philosophical Review* 60, no. 2 (April 1951): 177-97.

(iv) the judge feels his judgment is certain

(v) other competent judges render the same judgment in similar cases; and

(vi) the judgment is made "intuitively" rather than by "a conscious application of principles."

Rawls believes that when we attend to the judgments of those who have thought long and hard about moral matters and who are free of obvious bias, we often find that there is agreement, and he believes that this agreement justifies our adoption of their judgments and of the principles whose application coincides with them, the principles which, in Rawls' terminology, "explicate" them. Rawls writes,

> Since the principles explicate the considered judgments of competent judges, and since these judgments are more likely than any other judgments to represent the mature convictions of competent men as they have been worked out under the most favorable existing conditions, the invariant in what we call "moral insight," if it exists, is more likely to be approximated by the principles of a successful explication than by principles which a man might fashion out of his own head. Individual predilections will tend to be canceled out once the explication has included judgments of many persons made on a wide variety of cases.[44]

Perhaps the fact that we seem able to correct our distorted intuitions by interacting with others and examining an issue from different angles, as both Nagel and Rawls affirm, points to the possibility that we do have access to objective moral truth.[45]

[44] Ibid., 187.

[45] Rawls gives us an example of some judgments on which he believes the opinions of competent judges would converge: his well-known "principles of justice," defended at length in *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971).

While in the end I do agree with Nagel and with intuitionists that we have access to judgment-independent normative truths (Rawls doesn't declare himself on this question), and while I do agree that personal interests can lead us to have mistaken beliefs about them and that eliminating the influence of bias can result in convergence of our beliefs, I do not think that the fact that such a story of intuitive access to moral facts is *consistent* with our observations of moral disagreement automatically makes it the best explanation of our moral beliefs.[46] Another possible explanation—an antirealist one—is that all of our normative judgments arise from our personal interests, sympathies, and external pressures, and never from any connection to objective moral facts. If there is coming to be more and more moral agreement in the world, this may be not because we all share some core moral "knowledge," but simply because we are communicating more and more with one another, our cultures are becoming more and more homogeneous, and thus the interests we have and the social pressures we face are becoming increasingly similar. We ought to expect that the more our lifestyles converge, the more our attitudes will resemble each other's, but this explanation of moral convergence requires no appeal to some metaphysically unique sort of fact which influences our beliefs.

[46] Nagel does not believe this either. He provides a further argument for the objectivity of moral facts which appeals to the goodness of pleasure and the badness of pain (*The View from Nowhere*, 156-62). I won't critique this argument since I think it is essentially right. I do, however, disagree with Nagel's belief that we have special insight into the intrinsic moral properties of things besides phenomenal states.

Why, if intuitionists haven't been able to fill in any details about the nature of moral facts' influence on human minds, should one be inclined to accept their mysterious explanation rather than the antirealist explanation which is constructed entirely of facts about human attitudes open to scientific investigation? The intuitionist cannot simply appeal to the way that moral beliefs just strike them as reflecting judgment-independent truths, because, as we've already discussed, different people have feelings of self-evidence about different propositions that can't *all* reflect judgment-independent truths. But if self-evidence in *some* cases is not an indication that one has apprehended a judgment-independent moral truth, what reason do we have to believe that it is *ever* an indication of this? Even when we are sure that we are disinterested and impartial, and even when our moral judgments converge with those of others, the feeling of self-evidence that accompanies them may have a very different cause, perhaps one of those proposed by antirealists.

Now I'm not going to claim that intuitionists are entirely unjustified in holding onto the view that they have some access to objective moral facts. I think their feeling of conviction about realism may spring from sources which are in fact reliable but which they have not yet recognized. Nevertheless, I think that intuitionists should be deeply unsatisfied with the present incompleteness of their metaphysics and epistemology. If they want any chance of wooing antirealists to their point of view, or preventing a further flow of opinion toward antirealism, they are going to have to produce a much more substantial account of the nature of judgment-independent moral facts and of the relation they bear to our moral judgments, a relation which justifies us in believing that the latter (at least sometimes) reflect the former.

II. Minimal realism

Intuitionism's failure to produce an adequate picture of the relation between our moral judgments and the truths they supposedly reflect has led some realists to deny the need for any such picture, that is, to endorse a realism that is intentionally light on metaphysics, especially on anything that could be construed as a causal connection between moral facts and our moral judgments. A prominent proponent of this "minimal realism" is Ronald Dworkin.[47] According to Dworkin, metaphysical questions are just not the right sorts of questions to ask about ethics. Ethics, like mathematics, is its own domain of knowledge. Normative questions are not questions that can be answered by discovering anything that is "out there" in the world. (Dworkin coins the term "morons" to refer to the elementary particles of morality he believes it's ridiculous to postulate.) Normative questions can only be given normative answers. Goodness and badness do not causally affect our brains, just as numbers do

[47] T. M. Scanlon also seems to be a minimal realist, as he purposely dismisses the relevance of metaphysical questions about moral facts on pp. 2-3 of the introduction to *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press, 1998). Some other realists, though they do not completely dismiss metaphysical questions about moral facts, nevertheless do deny that any sort of causal connection between moral facts and our moral judgments is necessary to justify our belief in the objective truth of the latter. These realists include Shafer-Landau, Nagel (see, for example, pp. 144 and 148 of *The View from Nowhere* and p. 101 of *The Last Word* [New York: Oxford University Press, 1997]), Derek Parfit, and David Enoch. If, by denying the need for a causal connection, these philosophers mean to deny the need for *any influence at all* of moral facts on the formation of our moral judgments, then they, too, come under the purview of this section's criticisms.

not. But just as there are still truths (and falsehoods) about numbers, there are still truths (and falsehoods) about what is good and bad.

Dworkin uses his minimal realism to argue that widespread disagreement about ethical matters is no evidence against moral realism. He writes, "Whether diversity of opinion in some intellectual domain has skeptical implications depends on a further philosophical question: it has such implications only if the best account of the content of that domain explains why it should."[48] Because scientific thought, for example, is best explained by its connection to observable properties, wild disagreement about the properties of some object supposedly observed would make one suspicious that any object was actually observed at all. Dworkin offers the example of millions of people claiming to have seen unicorns and yet widely differing in their reports of the unicorns' appearance. He says we would feel justified in concluding that there had actually been no unicorns on the basis of the wide disagreement. "But," Dworkin continues,

> when we have no such domain-specific account of why diversity of opinion impeaches all opinion, we draw no skeptical conclusions from that diversity. Since we do not think that philosophical opinions are caused by philosophical facts, we do not conclude from the diversity of philosophical views (which is more pronounced than moral disagreement) that no positive philosophical thesis is sound.[49]

[48] Ronald Dworkin, "Objectivity and Truth: You'd Better Believe It," *Philosophy and Public Affairs* 25, no. 2 (Spring 1996): 87-139, p. 113.
[49] Ibid., 113-14.

The idea seems to be that, if we decline to accept the hypothesis that moral beliefs are caused by moral facts, then moral disagreement can't be taken as evidence that there are no moral facts. That is, if we reject the idea of a causal connection between moral facts and moral beliefs, it stops seeming suspicious that our moral beliefs differ. Moral disagreement may debunk a theory that says moral facts are observable, a theory which appeals to "morons," for example. Since it would seem hard for us to disagree about moral questions if moral particles were constantly staring us in the face, the existence of disagreement seems to count against the theory of morons. But Dworkin wants to say the existence of disagreement leaves untouched alternative theories of moral truth which do not postulate any causal connection between moral facts and moral beliefs.

The crucial question, however, is whether a theory that postulates judgment-independent moral truths can be plausible without defending some kind of interaction between those moral truths and moral judgments. Dworkin argues that it can. In the section of his article "Objectivity and Truth" specifically dedicated to questions of epistemology, Dworkin draws a distinction between beliefs about causally efficacious objects and events, and beliefs about other subjects. He says, "Since [beliefs about the physical world] are beliefs about objects and events that can interact causally with the human nervous system, it is sensible to include some requirement of direct or remote or at least potential interaction among our tests of their reliability. But nothing in the content of moral (or aesthetic or mathematical or philosophical) opinions invites or justifies such a test."[50] On the next page, he reiterates this contrast: "Since astrology and orthodox religion, at least as commonly understood, purport to offer causal explanations they fall within the large intellectual domain of science, and so are subject to causal tests of reliability. Since morality and the other evaluative domains make no causal claims, however, such tests can play no role in any plausible test for them."[51]

Dworkin seems to be saying that, because moral claims are not claims about the causal powers of moral facts, showing that there don't seem to be any causally efficacious moral facts cannot be evidence against the truth of these claims. That is true: moral facts could certainly exist without having any causal influence on anything. *However*, if one were able to show that moral facts were not able to influence our moral judgments in any way, causal or otherwise (and that moral facts and moral judgments were not both under the common influence of some third thing[52]), while this wouldn't provide positive evidence that our moral judgments were *false*, it would nevertheless defeat any justification we might have for taking them to be *evidence* for judgment-independent moral facts. *If our moral beliefs are not influenced by judgment-independent moral facts, then we would have the very same*

[50] Ibid., 119.

[51] Ibid., 120.

[52] I will leave this parenthetical clause implicit in the future.

moral beliefs whether or not judgment-independent moral facts existed, and thus our having these moral beliefs provides no evidence at all for their existence.[53]

Dworkin actually goes so far as to say that the nature of the content of moral beliefs excludes investigating whether they are influenced by moral facts. He says that we can't test the counterfactual claim "that the belief would not have occurred if the alleged cause had not been present…with respect to moral or aesthetic beliefs because we cannot imagine a world that is exactly like this one except that in that world slavery is just or *The Marriage of Figaro* is trash."[54] But while *we* may not be able to imagine such a world, the fact that others disagree or have disagreed with these judgments seems to show that they *do* imagine such a world. And the challenge (and opportunity) their disagreement presents is to determine *how* they have come to have this differing belief and whether the differences between their process of belief formation and our own reveal that one of us rather than the other has a better chance of reflecting a judgment-independent fact of the matter.

As we've already seen, Dworkin attempts to support his insistence that an investigation into the causal origins of our moral beliefs is unnecessary to justify belief in their objectivity by appealing to the fact that we do not believe such an investigation is necessary in the case of mathematical or philosophical belief. This appeal to an

[53] Realism's seeming inability to show that our moral judgments track judgment-independent moral facts in the face of evolutionary influences on our moral judgments forms the basis of Street's critique of realism in "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127 (2006): 109-66.
[54] Dworkin, "Objectivity and Truth," 119.

analogy with mathematics and non-moral philosophy is highly problematic. Realism about mathematical claims and the claims of non-moral philosophy is far from universally accepted. In fact, there are good reasons to be skeptical about realism in mathematics and in many philosophical domains. These reasons include protracted disagreement over many mathematical and philosophical questions,[55] as well as our inability to see how judgment-independent mathematical or philosophical facts could regulate (even imperfectly) our acceptance of them. That is, the very same questions that plague moral realism also plague realism in these domains.

Dworkin does not seem worried about antirealism in mathematics or in non-moral philosophy, however, and he doesn’t seem to think his readers will be worried, either. Why might this be? It does seem that in the case of mathematics, at least, there is a large number of its truths that in an obvious sense none of us doubts. We all know that employing mathematics produces reliable results not only in science and engineering but in our everyday lives. For instance, when we want to know whether a fleet of cars has enough seatbelts to transport a group of people, we’re quite confident that, if we correctly add up all of the seatbelts in all of the cars and see that this number is greater than or equal to the number of people in the group, then when

[55] For discussions of mathematical disagreement, see Hartry Field, “Which Undecidable Mathematical Sentences Have Determinate Truth Values?", in H. Garth Dales and Gianluigi Oliveri, eds., *Truth in Mathematics* (Oxford: Clarendon Press, 1998); and Penelope Maddy, *Naturalism in Mathematics* (Oxford: Clarendon Press, 1997).

everyone gets into a car, there will be enough seatbelts for all. For just such reasons as this, none of us is seriously tempted to deny claims such as "5 + 5 = 10."

But the utility of employing mathematical language and operations is not the issue in the debate over mathematical realism. What is in question is whether numbers and their relations *exist*, whether mathematical sentences are true or false in more than a pragmatic sense. If mathematics is to give us good reason to think realism in ethics is plausible, it's going to have to be because, in addition to its being clear that mathematical language and operations are useful, it's clear that numbers and their relations *exist* and serve as the ultimate truth-makers of mathematical sentences. But this is not at all clear; it's actually the subject of much debate in philosophy of mathematics.[56] And in the end, the same reasons that make it implausible to believe that our moral beliefs reflect judgment-independent moral facts that don't influence their formation make it implausible to believe that our mathematical beliefs reflect judgment-independent mathematical facts that don't influence them. If our beliefs are not somehow influenced by judgment-independent facts—i.e., if we would have the

[56] See, for instance, Hartry Field, *Science Without Numbers* (Princeton: Princeton University Press, 1980) and *Realism, Mathematics, and Modality* (Oxford: Blackwell, 1989); Philip Kitcher, *The Nature of Mathematical Knowledge* (Oxford: Oxford University Press, 1983); Penelope Maddy, *Realism in Mathematics* (Oxford: Clarendon Press, 1990); Stewart Shapiro, *Philosophy of Mathematics: Structure and Ontology* (New York: Oxford University Press, 1997); and Kit Fine, "The Question of Realism," *Philosopher's Imprint* 1, No. 1 (2001): 1-30.

very same beliefs whether or not judgment-independent facts on this matter existed—then our beliefs provide no evidence at all for their existence.[57]

Of course, there is an important way in which the debates over moral and mathematical realism differ. A lack of evidence for judgment-independent mathematical facts would give us no reason not to continue our employment of mathematical language and operations, since we do have evidence for the usefulness of this practice. Coming to antirealist conclusions about mathematics would thus have little to no effect on our daily lives (unless, of course, we're philosophers of mathematics with a strong personal interest in one conclusion or the other). On the other hand, a lack of evidence for judgment-independent moral facts may have the important consequences for our motivation and behavior that I described in Chapter 1. We seem to care more about the judgment-independent existence of moral properties than we do about the judgment-independent existence of numbers and their relations. (And this difference in concern may very well be justified by the very different natures of these two things.) Thus the fact that many of us tend not to worry ourselves about metaphysical questions in mathematics is no reason to conclude that we shouldn't worry about them in ethics.

[57] For similar arguments given specifically about mathematical realism, see Hartry Field, "Introduction: Fictionalism, Epistemology, and Modality," in his *Realism, Mathematics, and Modality* (Oxford: Blackwell, 1989); and Paul Benacerraf, "Mathematical Truth," *Journal of Philosophy* 70, No. 19 (1973): 661-79.

But if the lack of practical implications for our behavior stemming from the mathematical realism debate is a good reason not to extend our lack of worry about metaphysical questions in mathematics to metaphysical questions in ethics, what about the fact that we tend to think that there are judgment-independent answers to non-moral philosophical questions, despite our lacking a story about how judgment-independent philosophical facts influence our beliefs? The general philosophical case does seem to be a better analogy to the moral case than the mathematical one, both because there is even more evident disagreement on philosophical questions than on mathematical ones, and because the answers to such questions seem to have more potential for influencing our behavior. But does our lack of worry about widespread disagreement in other philosophical domains—when we have no metaphysical story explaining how some of us might still be getting at an objective truth of the matter—justify a similar attitude in the moral case, as Dworkin argues, or is it the other way around? Could it be that seemingly intractable disagreement over certain philosophical questions—without a metaphysical story to explain it—*is* a reason to abandon belief in a judgment-independent truth of the matter, or at least to abandon our pretensions to being able to discover it? We are back to our discussion of Shafer-Landau's philosopher and her justification for believing that her view about free will corresponds to a fact of the matter despite her recognition that others as intelligent as she, with all of the same evidence and the same amount of effort expended, have come to contrary conclusions.

I believe that philosophical questions in domains other than ethics *should* be submitted to an investigation of their potential for getting at a judgment-independent fact of the matter. I am not convinced that we have a non-undermining explanation for the existing wide divergence in beliefs about free will, the existence of universals, or the nature of reference, for example. It seems quite possible that the best explanation we have is that our beliefs on these subjects are not influenced by any judgment-independent fact of the matter, either directly or through other facts which are reliable indications of it. If these epistemological questions have as yet not been taken as seriously in these domains as they have in ethics, I think this must be largely due to the fact that these other philosophical questions, like questions of mathematical realism, have fewer and less important consequences for our everyday life than do ethical ones. In any case, it is not true that, because there is even more unexplained disagreement in philosophy in general than in ethics in particular, we are justified in not worrying about the disagreement that exists in ethics.

In the end, Dworkin doesn't exactly say we are justified in not worrying about the disagreement that exists in ethics. At some points in his article, he does seem to be concerned about the existence of disagreement and to think that, in the absence of objective reasons to think we are more likely to be getting things right than others, we should be "more modest" in regard to our moral beliefs.[58] But it's unclear why we should have any confidence at all in our moral beliefs if we think that their formation

[58] Dworkin, "Objectivity and Truth," 121-22.

is not influenced by the facts that give them their truth values. Dworkin suggests that, “[i]f you can’t help believing something, steadily and wholeheartedly, you’d better believe it.”[59] His strong sense of conviction about his ethical judgments seems to be what keeps him a realist despite the metaphysical and epistemological mysteries of his view (although he does say that not a great deal turns on the label “realist,” “given the notorious ambiguity of the term”[60]). He says, “it is startlingly counterintuitive to think there is nothing wrong with genocide or slavery or torturing a baby for fun. I would need very powerful, indeed unanswerable, reasons for accepting this….”[61] Certainly most of us will heartily approve of such conviction, and I don’t want Dworkin to abandon realism in spite of it. What I think he ought to do, however, is be concerned about finding a metaphysics and an epistemology adequate to support his conviction, if only because their lack persuades so many other people of antirealism.

Dworkin believes that his view already offers the most robust realism possible. He says, “there seems no more point in calling the view I have been defending minimalist than in calling it maximalist, because there is no more robust thesis for any realism to deploy or any anti-realism to refute, no more metaphysical a metaethics for the former to embrace or the latter to mock.”[62] And yet he doesn’t say this because he

[59] Ibid., 118.
[60] Ibid., 127.
[61] Ibid., 118.
[62] Ibid., 128.

can't imagine any more robust theories. He describes what he calls the "moral-field thesis" thus:

> The idea of a direct impact between moral properties and human beings supposes that the universe houses, among its numerous particles of energy and matter, some special particles—morons—whose energy and momentum establish fields that at once constitute the morality or immorality, or virtue or vice, of particular human acts and institutions and also interact in some way with human nervous systems so as to make people aware of the morality or immorality or of the virtue or vice.[63]

But Dworkin concludes that if this view "is intelligible, it is also false. It is not even a remotely plausible thesis to attribute to anyone…quite apart from its insanity as a piece of physics."[64] Elsewhere he calls it "mysterious," "artificial," and "counterintuitive."[65] No doubt the particular theory that he describes is a bit bizarre, but two things can be said in response to this. Firstly, this theory of moral atoms does not represent the only way in which moral properties could influence our moral beliefs. In the next chapter, I will present a more compelling picture of the place of moral properties in the empirical world. Secondly, if one is setting out to determine how moral properties could influence our beliefs, one should expect a somewhat extraordinary answer. The nature of judgment-independent moral facts has been a puzzle for too long for its solution not to end up challenging existing conceptions of the world.

[63] Ibid., 104.
[64] Ibid., 104-5.
[65] Ibid., 118.

In the end, I think that minimal realism does reflect many people's minimal understanding of the status and justification of their moral judgments, and it may be useful as a last resort against antirealism. But I don't think it's a position that will sustain philosophers for very long. I think it is ultimately a way-station on the path to antirealism or to a metaphysically and epistemologically robust realism.

III. Ideal-observer theory

Neither intuitionism nor minimal realism seems to have the resources to make plausible the claim that we sometimes have knowledge about judgment-independent moral facts. In particular, these theories have not been able to offer—or have flatly refused to offer—an account of the relation between moral judgments and moral facts which explains moral disagreement in a more plausible way than antirealism. There are two other types of self-described realist theory, however, that make it their goal to succeed where intuitionism and moral realism fail. These are ideal-observer theories and theories based on a synthetic ethical naturalism.[66] Both of these types of theory focus on providing an epistemology which explains both how moral facts are epistemically accessible to us and why we are often mistaken about them, and thus often disagree with one another on moral questions. Unfortunately, these theories have their own set of problems. Either their account of value is not ultimately judgment-

[66] Though I discuss these theories separately, it is possible that a single theory could fall into both of these categories.

independent, or their account of value *is* judgment-independent, but it is either arbitrary and unconvincing, or incomplete.

I'll begin by discussing ideal-observer theories. Their prominent proponents include Roderick Firth,[67] Richard Brandt,[68] Peter Railton,[69] and Michael Smith.[70] Ideal-observer theories have evolved a fair amount between Firth's outline of the view in 1952 and the versions currently defended by Railton and Smith. What all ideal-observer theories have had in common, however, is an account of value which appeals to what someone would desire or take to be valuable were he an "ideal observer," i.e., were he to possess such things as perfect knowledge of the world and/or a perfect ability to reason. In my discussion here, I am going to focus on the particular view which Railton has developed, but my criticisms are, with slight modification, applicable to other versions of ideal-observer theory as well.

Railton appeals to the notion of an ideal observer to provide an account of an individual's "objective interests." According to Railton, what is objectively in a

[67] Roderick Firth, "Ethical Absolutism and the Ideal Observer," *Philosophy and Phenomenological Research* 12, No. 3 (March 1952): 317-45.
[68] Richard B. Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979).
[69] Railton, "Moral Realism"; "Facts and Values," *Philosophical Topics* 14 (Fall 1986): 5-31; and "Naturalism and Prescriptivity," *Social Philosophy and Policy* 7, no. 1 (Autumn 1989): 151-74.
[70] Michael Smith, "Moral Realism," in Hugh LaFollette, ed., *Blackwell Guide to Ethical Theory* (Oxford: Blackwell, 2000), 15-37; *The Moral Problem* (Oxford: Blackwell, 1994); "Reason and Desire," *Proceedings of the Aristotelian Society* 88 (1987-88): 243-56; "Dispositional Theories of Value," *Supplement to the Proceedings of the Aristotelian Society* 63 (1989): 89-111; "Response-Dependence Without Reduction," *European Review of Philosophy*, Special Issue on Response-Dependence, eds. Roberto Casati and Christine Tappolet, 3 (1998): 85-108; "Does the Evaluative Supervene on the Natural?", in Roger Crisp and Brad Hooker, eds., *Well-Being and Morality: Essays in Honour of James Griffin* (Oxford: Oxford University Press, 2000), 91-114; and "Exploring the Implications of the Dispositional Theory of Value," *Philosophical Issues: Realism and Relativism* 12 (2002): 329-47.

particular individual’s interests is what that individual—were he endowed with perfect factual and counterfactual knowledge of the world and of himself, as well as with unqualified cognitive and imaginative powers—would want his non-idealized self to want. For each of us actually existing persons, our objective good is equivalent to what our idealized self would want us to want in our actual situation. Railton then uses this account of individuals’ objective interests to construct an account of morality. He defines moral rightness as “what would be rationally approved of were the interests of all potentially affected individuals counted equally under circumstances of full and vivid information.”[71]

On Railton’s view, and on ideal-observer theories generally, the great gap between the perfection of the ideal observer and our actual knowledge and reasoning ability is meant to account for our frequent mistakes (and hence disagreements) about what is truly valuable. The idea is that we often disagree about what is good or right simply because our desires are not in line with what they would be were we better informed and better able to follow through every complicated deduction or probabilistic inference. Railton provides the example of two dehydrated travelers, one who craves clear liquids and the other who craves milk.[72] The one who drinks the clear liquids quickly feels better, while the one who drinks milk does not. It seems that this is an obvious case in which one person’s desires were in line with his objective

[71] Railton, “Moral Realism,” 190.
[72] Ibid., 174-5, 178-80.

good while another's were not. And had the traveler who craved milk known how much better he would have felt after drinking clear liquids, he would no doubt have desired to drink clear liquids instead. His mistake about his own good was based on a lack of information. According to ideal-observer theory, we often make mistakes about what is good for the very same reason we make other mistakes—for lack of information and brain power—and, given this explanation, there is no reason to conclude from the existence of disagreement over what is valuable that there is no fact of the matter.

There is an important further question to ask, however, and that is whether the "fact of the matter" about what is valuable that is appealed to by ideal-observer theories is realist in the sense I have defined: i.e., whether it is judgment-independent. The gap between what an ideal version of an individual would desire and what that individual actually desires is what Railton cites as giving his account of value "objectivity." That which an idealized version of someone would desire is supposed to deserve the label "objectively good" because it does not depend on what he actually desires in less-than-ideal conditions. However, while it seems clear that certain of our desires would change if we were to become ideal observers, I am going to argue that there are other desires of ours that becoming an ideal observer would not change at all, and in fact that these desires would control all of the other changes in our desires. If this is so, then what is supposedly "objectively" good for an individual according to

Railton's ideal-observer theory is actually highly dependent on his or her *actual* desires, and this is a major threat to his theory's status as realist.

Railton provides us with the following description of the phenomenon of desire with which his theory is concerned.

> Consider first the notion of someone's *subjective interests*—his wants or desires, conscious or unconscious. Subjective interest can be seen as a secondary quality, akin to taste. For me to take a subjective interest in something is to say that it has a positive *valence* for me, that is, that in ordinary circumstances it excites a positive attitude or inclination (not necessarily conscious) in me.[73]

The only difference between this sort of desiring, which constitutes one's "subjective" interests, and the desiring that constitutes one's "objective" interests, according to Railton, is that the latter is what would occur under conditions of perfect knowledge and infinite cognitive ability.

Now one can see why coming to have full information about oneself and one's environment and having unlimited cognitive abilities could produce a significant change in one's attitudes and inclinations. As in the case of the traveler who would drink clear liquids rather than milk if he knew what the effects on him of each would be, one would no doubt discover that many of the things one was previously inclined to do would actually have effects one dislikes, and that certain other courses of action would bring about results one prefers. Once one came to believe the truth about the consequences of all of one's possible actions—and to be able to imagine these

[73] Ibid., 173.

consequences vividly—one’s attitudes and inclinations would naturally shift in favor of the actions whose consequences one preferred. However, note that such a shift in attitude about one’s actions would take place as a result of one’s further, deeper preferences about their ultimate consequences, preferences which need *not* have been affected by increased knowledge and increased cognitive abilities. It seems that any change in desires prompted by increased knowledge or increased cognitive abilities will depend on a deeper desire—a desire about ultimate consequences—which does not go through any process of modification and thus not through any process of objectification.

Some ideal observer theorists will no doubt resist this claim, asserting instead that fuller and more vivid knowledge of the world *can* affect our ultimate desires and is not limited merely to improving instrumental ones. I am thinking particularly of Brandt, who holds a fairly complex view about how “intrinsic desires,” which he defines as “ones not obtaining because the wanted event is thought to be a means to satisfaction of another desire,” can still undergo modification in the presence of certain sorts of knowledge, such as knowledge about their origins.[74] Brandt suggests several possible facts about the origins of our desires that, if we were to come to know them, would tend to “extinguish” those desires. He claims that we will tend to reject desires that we come to believe have been acquired solely for one or more of the following

[74] Brandt, “The Science of Man and Wide Reflective Equilibrium,” *Ethics* 100, no. 2 (Jan. 1990): 259-78, p. 262. See also his *A Theory of the Good and the Right*, Ch. 6.

reasons: (1) we were in a state of satiation, deprivation, elation, or depression, (2) we associated an object or situation with the approval or disapproval of another party, or (3) we generalized from an untypical example (for instance, we ended up associating with trauma an object or situation very unlikely to produce such trauma). To use one of Brandt's frequent examples, if we came to believe that we desired some object only because of some off-hand remarks once made by our second-grade teacher, we would likely give up the desire, or rather, we would come to find it "extinguished" by this belief.

Confrontation with the origins of our desires is in fact one of the tools psychotherapists successfully use to combat irrational fears and other harmful behaviors, and I don't dispute its effectiveness. (Brandt appeals to the results of psychotherapy to support his view.) I do not, however, think that the cases Brandt describes are ones in which desires are modified for non-instrumental reasons. Each time that a desire is brought up for consideration, what is considered is whether it is a "sensitive response to the real world," to use Brandt's phrase.[75] Desires are extinguished to the extent that they are thought to be the result of arbitrary associations of objects or situations with positive or negative consequences or feelings, consequences or feelings to which they have no necessary or even probable contingent connection, apart from the isolated occasions on which they were formed. What this test evaluates are the *actual consequences* of the pursuit of an object, and

[75] Brandt, "Science of Man," 262, 275.

when these are thought not to justify one's desire for the object, one's desire tends to extinguish.

But this means that these desires, described by Brandt as non-instrumental, are actually being evaluated, and then discarded or retained, based on one's desires vis-à-vis their consequences. This does not mean that these desires, which Brandt calls "intrinsic," were ever consciously endorsed because of a belief that they would satisfy further desires. In this sense, perhaps, they could be called "ultimate" desires. But they are not ultimate in the sense with which I am concerned above, because they are in fact subject to *change* based on other desires about their consequences. Thus my claim that "any change in desires prompted by increased knowledge or increased cognitive abilities will depend on a deeper desire—a desire about ultimate consequences—which does not go through any process of modification" is entirely consistent with Brandt's actual examples, and is even supported by them, since any change in desires of the sort described by Brandt requires a further desire which evaluates the consequences of the first as negative.

My claim is also meant to be consistent with the belief that some changes in our desires occur without being bidden by a deeper desire. It seems that all sorts of things could cause one's desires to change without reference to a further desire. Perhaps, given the right wiring in one's brain, any sort of new information could causally produce any sort of change in one's most basic desires. However, Railton and Brandt are not going to want to appeal to this sort of change in desires, since the purpose of appealing to the modified desires of an ideal observer is to appeal to desires which are *less arbitrary*—and thus seemingly more “objective”—than those one actually has. Railton and Brandt should want to limit changes in an ideal observer’s desires to cases in which there is some *reason* for his desires to change in view of the new information and capacities he acquires. But it seems there can only be such a reason if (1) he has a further desire with which to evaluate the consequences of the first, or (2) the new information he acquires includes information about a moral fact.

The second option is a possibility which Railton doesn’t discuss, presumably because he wants ideal-observer theory to explain moral facts rather than appeal to them. (I will nevertheless discuss this possible development of ideal-observer theory at the end of this section.) Assuming that Railton doesn’t want his ideal-observer theory simply to appeal to moral facts, he is going to have to accept that any rational change in our desires based on the new information we gain in becoming ideal observers is only going to take place with the help of some further desire from the perspective of which we evaluate that new information. But this means that there are going to be some desires of ours which no amount of new information will give us reason to change, as they are themselves the ultimate basis for any such reasons.[76] And if our most basic desires about ultimate consequences will remain untouched by the

[76] One might try to appeal to a non-hierarchical set of desires, such that each desire could form a rational basis for changing another and none is ultimately immune to change. But if none of our desires are deeper than others, it seems we will be in a position of never being able to make a non-arbitrary choice about which desire to give up when two or more come into conflict. A completely non-hierarchical set of desires does not provide a basis for rational change in desire.

prodigious new mental powers that transform us into ideal observers, and if these desires in fact control the changes in all of our other desires, then our “objective” interests on Railton’s theory are actually highly dependent on certain of our actual desires. Railton looks like he’s not ultimately offering a judgment-independent account of moral facts. His theory looks very much like a version of constructivist antirealism.[77]

Whether Railton’s theory *is* ultimately antirealist depends on the status of its most central claim: “What is good for one is what one’s idealized self would desire one to desire.” Railton could intend this claim to be true by definition, simply as a statement of what we *mean* when we say something is good for someone. If it’s taken this way, then the view is antirealist: there are no moral facts that don’t simply reduce to facts about what one actually desires (because of the way we’ve seen that facts about what one’s idealized self would desire reduce to facts about what one actually desires). If this claim is instead taken to be substantive, then we have to ask the further question: Is the truth of *this* claim independent of all of our normative judgments? If Railton says “no,” then his view is again antirealist.[78] If he says “yes,” however, then he has the further task of explaining what sort of objective moral fact this claim reflects and how we could come to know its truth. That is, if Railton chooses the

[77] For an example of a philosopher who categorizes Railton’s theory as antirealist, see Street, “A Darwinian Dilemma for Realist Theories of Value,” 137.

[78] For a similar conclusion, see Street, 137. Discussing Railton’s view as a version of what she calls “value naturalism,” she says, “In order to count as genuinely realist, … a version of value naturalism must take the view that *which* natural facts evaluative facts are identical with is independent of our evaluative attitudes.”

realist route, he will ultimately have to appeal to some other version of realism to provide epistemological support for this central tenet of ideal-observer theory, and will confront the same problems other realisms do.

But in addition to facing the problems of other realisms, Railton's view has its own particular implausibility, due to the potential (and often actual) arbitrariness of desires, including those which determine our "objective" interests. Our most basic desires, those which would not change were we to become ideal observers, are, like all of the others, defined by Railton as "positive attitudes or inclinations," though they are very foundational attitudes or inclinations, ones which serve as a standard of evaluation and justification for all of our other, derivative ones. The question arises: Would it not be theoretically possible for us to have a foundational positive attitude or inclination toward *anything*? Aren't the particular ultimate goals which prompt us to action simply a function of the wiring of our brains, which in turn is a contingent result of our evolutionary and personal history? Is it not a contingent matter that we are generally motivated by our own health and safety, by pleasure and the avoidance of pain, by any of the things that we as a matter of fact desire as ultimate ends? Or, if it's impossible for us not to have a desire for some of these things—if, for instance, some desire for pleasure and the avoidance of pain is too deeply integral to the functioning of our brain—couldn't we nevertheless have developed all sorts of additional ultimate desires, if our evolution had taken a different course? This seems highly likely, given that there are plenty of animals which appear to have ultimate desires which differ substantially from ours, and given that there are many human beings who, for one reason or another, have different ultimate preferences. It also seems that we are theoretically capable of changing people's ultimate desires by rewiring their brains to give them positive attitudes toward different objects.

But if our ultimate desires are so highly contingent and malleable, what reason do we have for thinking that they are the basis for anything deserving the name of "objective value"? It doesn't seem that there needs to be anything good about the objects of our desires *themselves* in order for us to desire them as ultimate ends. It suffices for our brains to be wired in certain ways. It thus seems quite odd to call whatever one happens to have a foundational positive attitude or inclination toward "objectively valuable."

Railton might reply by pointing out that the sort of value that is constituted by something's being the object of desire shouldn't be understood as something's being valuable *in itself, for everyone*. One of the virtues of ideal-observer theory is that it recognizes that different things can be valuable for different people, something that certainly seems true. According to Railton, an individual's desires determine what is valuable, not for everyone, but for *that individual*. If someone's ideal self would desire something for him, then that is what is objectively valuable for *him*.

Yet even if it's acknowledged that an individual's ultimate desires only determine what is objectively valuable for that individual, it can still seem quite odd that, since any individual's desires could be rewired, any of a multitude of things could be objectively good for him, depending on what he happens to be wired to want at a particular time. For example, any of us could plausibly be rewired in such a way as to be disposed to seek out extreme bodily harm or death as our ultimate goal, even if we were healthy individuals living painless lives. Railton would have to say that, in such a case, extreme bodily harm or death would be in our objective interest.

Simply because almost any desire could be induced in almost any individual (given that they have the ability to conceptualize or discriminate its object), it seems that desire can't be enough to make its object objectively valuable, even for that individual. Whatever we happen to desire as an ultimate goal, we can ask the further question whether we *should* desire that thing. It seems entirely intelligible to ask what is *good* for us, regardless of what we may desire, and regardless of what ultimate desires we may retain even under conditions of full information and unlimited cognitive ability.

This divide between desire and objective value comes into even sharper relief when we consider the desires of others. Railton's theory depends on moral rightness being a construction out of the equal consideration of all individuals' objective interests: i.e., their desires under ideal conditions. But why, one might wonder, do we have any obligation to satisfy the desires of others? Is the fact that someone else is moved to seek out a certain thing any reason for us to seek to acquire that thing for them as well? The fact that another person ultimately desires bodily harm or death doesn't seem a sufficient reason for me to be morally obligated to advance these goals in any way. Others' desires will motivate *them* surely enough, but the fact that *they* are motivated in that way doesn't mean *I ought* to be so. I might be able to see how I have an obligation to promote something which is *objectively* good for others, but when this "objective" good is defined as whatever satisfies their ultimate desires, the normative force of the obligation seems to disappear. How can I be obligated to promote the satisfaction of something so arbitrary?

Perhaps it will be suggested that what makes the satisfaction of desires—both ours and others'—a non-arbitrary goal is the fact that satisfying desires produces pleasure or reduces pain. Don't the procurement of pleasure and the avoidance of pain seem non-arbitrary goals? In fact, I believe that the promotion of pleasure (broadly understood) and the avoidance of pain (also broadly understood) are the *only* non-arbitrary goals we could have, and that they form the basis for all of our objective normative reasons. But the ideal-observer theorist doesn't say that we ought to satisfy people's desires only insofar as such satisfaction would promote pleasure and avoid pain. The ideal-observer theorist says we ought to promote the satisfaction of *all* people's ultimate desires (i.e., all those desires that would remain if they became ideal observers), no matter what these desires are, and no matter whether their satisfaction best promotes the balance of pleasure over pain in their lives or not.

Now one might hold the view that everyone's ultimate desires are always simply to promote pleasure and avoid pain, but this is a very controversial empirical hypothesis, and I know of no ideal-observer theorist who claims this. If one did claim this (as John Stuart Mill did, for instance[79]), and defended it well, one could possibly make a good case for ideal-observer theory's having a non-arbitrary notion of objective goodness. But in that case, all of the non-arbitrariness of this notion of objective goodness would be derived from the non-arbitrariness of promoting pleasure and avoiding pain, not from any non-arbitrariness inherent in the idea of satisfaction of desire itself. In the end, I think it's fairly clear that we actually have ultimate desires for things besides pleasure and the avoidance of pain and in fact could have had any of a great variety of such non-hedonistic desires if evolution had seen fit to produce them in us, or if we rewired our neural circuitry in the appropriate ways. Thus to embrace Railton's theory is to advocate the satisfaction of potentially very arbitrary desires. The central tenet of his theory—"What is good for one is what one's idealized self would desire one to desire."—is an implausible account of judgment-independent good.

But there remains a way in which a plausible realist interpretation of ideal-observer theory might be achieved, as I previously mentioned very briefly. It might be asserted, contrary to what we have been assuming, that gaining complete knowledge of the world *would* affect an individual's ultimate desires. What if part of becoming an ideal observer was gaining knowledge *of what is valuable*? That is, what if one of the sets of facts a person could know about the world was a set of facts about what is objectively good for him? Coming to know what is objectively good for him might

[79] John Stuart Mill, *Utilitarianism* (1863).

conceivably cause him to desire that thing. If this were the case, then although desires could be arbitrarily programmed into someone, if that person became omniscient, he would then know what was truly good for him and his desires would change accordingly.

This view would have to be considerably more complicated than the ideal-observer theories currently advocated. There would be much further theory to add, explaining things such as how it is that knowledge can influence ultimate desires and whether it always does or under what conditions it does. I don't think developing this sort of theory would be impossible. I can't say that the sheer complexity of it should deter anyone, given the project I myself am undertaking in developing my own view. Yet the other thing that such a view would have to add is an account of objective, desire-independent value. It would have to explain what exactly these facts are that change one's desires when one comes to know them, and it would have to explain how one could come to know them. I actually think it's possible to give such an account, but if one has such a good handle on objective value independent of any reference to an ideal observer, then it seems one's theory is not truly an ideal-observer theory anymore. It is a robust realism, perhaps much like the one I am going to put forward.

In the end, it seems to me that existing ideal-observer theories are actually best classified as antirealist, because of the way they make goodness dependent on our actual ultimate desires (or on other actual attitudes or judgments of ours, depending on the version of ideal-observer theory). But, as we've seen, if an ideal-observer theorist wants to cling to realism, he has two options. First, he could explain what sort of objective moral fact the central tenet of ideal-observer theory reflects and how we could come to know it. This presents the ideal-observer theorist with the same metaphysical and epistemological problems that other realist theories have, and that ideal-observer theory was supposed to avoid. In addition, the moral fact whose existence the ideal-observer theorist then has to defend seems particularly implausible, given the potential, and often actual, arbitrariness of desires (an arbitrariness which also extends to any other evaluative attitudes or judgments of ours, if these are not understood to reflect any judgment-independent moral facts). The realist's second option is to claim that the ideal observer's ultimate desires would not be arbitrary because they would be influenced by his knowledge of objective good. But as we've just seen, explaining and defending such a picture would require having such a good handle on the nature of objective value that an appeal to the desires of an ideal observer would be superfluous. The view would already have solved the major metaphysical and epistemological puzzles of realism in some other way. In the end, it seems that ideal-observer theory is not the key to producing a plausible account of judgment-independent moral facts.

IV. Synthetic naturalism

The final sort of realism I am going to discuss is synthetic naturalism. Theories of this type tend to offer a more concrete description of objective value than do ideal-observer theories. They claim that moral goodness is a natural property or a cluster of natural properties (perhaps related in a very complex way), and they claim that which natural property or properties are identical with moral goodness is an *empirical* question, answerable only through observation and investigation over time. Just what sort of observation and investigation is required, however, varies with the version of synthetic naturalism, and it is synthetic naturalism's inability to settle this question in a non-arbitrary manner that makes it inadequate as a realist theory.

Let's start by examining the view of Richard Boyd.[80] Boyd calls moral goodness a "homeostatic cluster property." According to Boyd, moral goodness is identical with the satisfaction of important human needs and the promotion of the homeostatic mechanisms which tend to make the things that satisfy them mutually supporting.[81] Boyd gives as examples of important human needs physical and medical needs, as well as "the need for love and friendship, the need to engage in cooperative efforts, the need to exercise control over one's own life, [and] the need for intellectual and artistic appreciation and expression."[82] Some of the mechanisms that he cites as contributing to the mutual satisfaction of these needs are psychological and social mechanisms such as "cultivated attitudes of mutual respect, political democracy, egalitarian social relations, various rituals, customs, and rules of courtesy, [and] ready

[80] Richard Boyd, "How to Be a Moral Realist," in Geoffrey Sayre-McCord, ed., *Essays on Moral Realism* (Ithaca, NY: Cornell University Press, 1988), 181-228.
[81] Ibid., 203.
[82] Ibid.

access to education and information."[83] But Boyd also mentions that the question of just which important needs human beings have is a difficult one whose answer requires extensive empirical investigation into human psychology and biology.

Boyd stresses, like all synthetic naturalists, that he is not giving an analytic definition of moral goodness but is opening the way for the discovery of a *synthetic* identity between moral goodness and some set of natural properties. Which set of natural properties is identical with moral goodness depends, according to Boyd, on which natural properties causally regulate our use of the term 'good'. He believes that the needs he has listed play an important role in this regulation, but that by further empirical investigation, we can discover other aspects of the homeostatic mechanisms which unify our use of the term 'good'. We come to realize that the satisfaction of certain needs is identical with moral goodness not because the satisfaction of these needs and moral goodness are identical by definition, which would be a highly implausible claim, but in the way that we have come to realize that water is identical with H_2O. Our concept of water does not include its being H_2O, but empirical investigation of water has led us to the conclusion that water and H_2O are nevertheless the same thing. In the same way, Boyd hopes that, though our concept of moral goodness does not explicitly include the satisfaction of all of the important human needs, empirical investigation will continue to reveal that our use of the term 'good' is causally regulated by a homeostatic cluster of things that satisfy these needs, the

[83] Ibid.

specific characters of which will become increasingly clear. And it will thus become increasingly clear just which objective, natural properties we have all along been referring to in talking about goodness.

The most prominent objection to appeals to synthetic identities between the referents of ethical terms and the referents of natural property terms is that raised by Terence Horgan and Mark Timmons in their paper "New-Wave Moral Realism Meets Moral Twin Earth."[84] Horgan and Timmons ask us to consider a planet much like Earth—call it "Moral Twin Earth"—where the inhabitants also use moral terms like 'good', 'bad', 'right', and 'wrong' and tend to carry out the actions they call "right" and avoid those they call "wrong," just as we tend to do on Earth. The difference between Moral Twin Earth and Earth is that their inhabitants' use of these terms is causally regulated by different natural properties. (Horgan and Timmons propose that moral language on Moral Twin Earth might be causally regulated by properties capturable in a deontological normative theory while moral language on Earth is causally regulated by properties capturable in a consequentialist normative theory.) The question, then, is whether the reference of the moral terms differs between Earth and Moral Twin Earth due to this difference in causal regulation. Do the terms refer to

[84] Terence Horgan and Mark Timmons, "New-Wave Moral Realism Meets Moral Twin Earth," *Journal of Philosophical Research* 16 (1991): 447-65. See also their "Troubles on Moral Twin Earth: Moral Queerness Revived," *Synthese* 92 (1992): 221-60, and Timmons's *Morality Without Foundations* (Oxford: Oxford University Press, 1999), Ch. 2.

one set of natural properties in the mouths of Earthlings, and another set in the mouths of Twin Earthlings?

The assumption is that Boyd, because he says that reference is determined by facts about causal regulation, will have to say that Earthlings' and Twin Earthlings' terms do refer to different properties. The problem in that case is that, if Earthlings and Twin Earthlings were to engage in a discussion involving the term 'good', they would not be able genuinely to disagree about what is good, but would only be talking past one another. If an Earthling said, "Satisfying need N is good," he would not be disagreeing with the Twin Earthling who said, "Satisfying need N is not good," since 'good' for the Earthling refers to one set of properties (which let's say does include satisfying N) while 'good' for the Twin Earthling refers to a different set of properties (which let's say does not include satisfying N). Yet Horgan and Timmons insist that it is much more natural to understand the Earthling and Twin Earthling as engaging in a real moral disagreement over what is good in some sense that they both understand. They seem to be arguing about what ought to be done, about whether it is morally important that one should satisfy need N. They seem to be engaged in a conversation which has as part of its purpose for each of them to persuade the other to act in a certain way in the future. Since this sort of disagreement and attempt at persuasion is not possible if the Earthling and Twin Earthling are not referring to some common property by their uses of 'good', but each referring to a different set of natural properties, Horgan and Timmons conclude that we should not understand the reference of moral terms as fixed by the causal regulation of their usage. Thus, in their minds, Boyd cannot help himself to a synthetic identity between goodness and the natural properties which causally regulate the use of the term ‘good’. However our current and historical use of the term ‘good’ has been causally regulated, it is an open question whether we are *correct* in this application of the term. Two people who have in the past called very different natural properties “good” can be understood as having disagreed with one another and thus do have the potential to come to an agreement.

I believe Horgan and Timmons are correct in their criticism of Boyd’s view. Our concept of goodness has a fuller meaning than is given by the way in which its use is actually causally regulated. And this fuller meaning makes it possible that many, or even most, of our applications of the term in the past have been mistaken.

Synthetic naturalists have attempted to deal with this criticism, however. David Brink, in a defense of the causal theory of reference, proposes that we could understand reference as determined not by *actual* causal regulation of a term’s use, but by its *counterfactual* causal regulation. On this proposal, what determines a term’s reference is not the ways in which it is employed in the actual world but the ways in which it would be employed “upon due reflection in imagined situations and thought experiments.”[85] Particularly, a term’s reference depends upon the natural properties that would regulate a person’s use of the term if his beliefs were in “dialectical

[85] David Brink, “Realism, Naturalism, and Moral Semantics,” *Social Philosophy & Policy* 18, no. 2 (Summer 2001): 154-76, p. 168.

equilibrium," where the proper understanding of dialectical equilibrium is "broad, representing a dialectical accommodation not simply among our moral beliefs, but among our moral beliefs and various philosophical and empirical beliefs."[86] If we understand reference as a function of this sort of counterfactual causal regulation, then we could possibly allow that Earthlings and Twin Earthlings are in disagreement over the proper application of the term 'good'. Their differing current and past usage of the term does not preclude the possibility that Earthlings and Twin Earthlings would nevertheless apply the term identically (or at least in an adequately similar manner) after reflection, and if their usage would indeed converge as their beliefs approach dialectical equilibrium, then they do currently disagree over the application of a term with a single reference that they are both in the process of discovering.

This is an ingenious way of patching up the causal theory of reference, but I think it still leaves us with an unsatisfactory realism. This innovation suggested (though not adopted) by Brink returns us to a variation on ideal-observer theory. We are told that certain natural properties constitute the referents of moral terms because an ideal "reflector" would apply the moral terms in all and only those cases where these natural properties are instantiated. And yet, just as in the traditional ideal-observer theories the desire of an ideal observer ultimately appears arbitrary unless it is founded on an independent account of moral goodness, so the usage of moral terms by an ideal reflector will ultimately appear arbitrary unless it is founded on an

[86] Ibid., 169.

independent account of moral goodness. Without such an independent account, we cannot know, even if there is convergence in the use of moral terms by ideally reflective speakers, that it is non-arbitrary. Individuals' usages of terms might converge for any number of reasons, including the fact that they were all raised in cultures where moral terms were used in similar ways. What is needed from a theory of moral realism is not merely a hypothesis of possible convergence in the application of moral terms. Even if it could be proved that everyone would eventually, after adequate reflection, apply moral terms in the very same way, it would still remain to be shown that this convergence was a result of goodness and badness' actually being objective properties instantiated in the world and not the result of some arbitrary similarity in our dispositions. Even an antirealist, for example, could believe that everyone who has reached dialectical equilibrium would express similar attitudes toward various actions and states of affairs and express these attitudes by saying that certain things are "good" and others "bad." And yet, for all that, the antirealist would not have to concede that there was anything intrinsically normative about the natural properties everyone converged on calling "good" and "bad."

To make a proposal like Brink's thoroughly realist, what is needed is an account of why there would be non-arbitrary convergence in the application of moral terms. The theory needs an account of *why* an ideal reflector would use moral terms in a particular way. When considering using the term 'good' in a particular situation, what criteria does the ideal reflector employ? She can't simply tell herself to use the term “however an ideal reflector would.” The ideal reflector herself, if she is not to be arbitrary, must possess some concept of goodness which she uses as her standard in evaluating any potential object of praise. But if such a concept exists, then it is that concept which ultimately gives meaning to the term ‘good’, and the question of realism becomes whether there is any judgment-independent connection between this concept and a property or properties instantiated in the world. The crucial element in a theory like the one Brink suggests would not be convergence in usage by ideally reflective speakers, but *non-arbitrary* convergence, convergence due to our recognition that certain things satisfy our concept of goodness. This in turn means that once again what is required is a more robust realist theory: one that provides an account of our concept of goodness, the way in which certain things objectively satisfy it, and the way in which we come to know this.

The proposal of Brink’s that I have been discussing is not one that he actually endorses. Brink himself prefers an account of reference based on *referential intentions*. On this account, the reference of a speaker’s terms depends on what the speaker herself intends. Brink suggests that we could understand Earthlings and Twin Earthlings’ disagreement in virtue of their matching intentions to pick out with the term ‘good’ “the properties—whatever they are—of people, actions, and institutions that make them interpersonally justifiable.”[87] I hope that it is clear how this proposal faces the same ultimate problems as the last. In order to be thoroughly realist, it will

[87] Brink, 175.

need to provide an account not just of what properties people *take* to be justifying but of what properties *are* justifying. Just as people's desires could be arbitrarily wired, so could their tendencies to accept certain justifications. Without an account of what properties are actually justifying, a theory like this is either antirealist or highly incomplete.

I believe that, in the end, every version of synthetic naturalism that aspires to realism is going to need an account of our concept of goodness (or of rightness, or of whatever it takes the basic moral concept to be) and of how we come to recognize that things in the world objectively satisfy it. In the case of verifying the synthetic identity between water and H_2O, there is no problem with nailing down the first half of the identity. We know how to find water in the world: we look for clear, colorless, tasteless liquid. Once we've found this, we can examine its molecular structure. In the case of goodness, however, the synthetic ethical naturalists can hardly claim to have nailed down the first half of the identity. The synthetic naturalists have not told us what the property of goodness looks like or feels like in the world. Nor have they given us any other non-arbitrary means of locating it, that is, a means of locating it which appeals in some way to our concept of goodness, rather than simply observing to what objects we happen to apply our term. But until we can identify the property of goodness in the world in some non-arbitrary way, we don't have the possibility of more deeply exploring its structure and discovering it to be identical with anything else. Thus before we can discover any synthetic identities involving goodness, we have to understand just what our concept of goodness is and whether it has any judgment-independent connection to things in the world.

V. Criteria for a plausible realism

In my discussion of current realist theories, I hope to have clearly brought out a theme: that current realist metaethics is lacking the basic metaphysical and epistemological framework necessary to make it a plausible alternative to antirealism. The problems we've seen in existing realist theories suggest a list of four basic elements that a realist theory ought to have in order to be a plausible alternative to antirealism. A satisfactory realist theory will, at a minimum, provide the following:

(1) An account of our concept of goodness.[88] When we use the word 'good', or when we merely contemplate goodness, what is it that we have in mind? If we are to know to what things in the world our concept of goodness objectively applies, we must have some handle on what this concept is. And after all, if we don't even know what we mean by 'goodness', what's the point of insisting that anything objectively has it?

We may not be able to give a non-circular definition of 'goodness', but this criterion does not require a definition per se. We simply need to be able to have before

[88] A theory could also be adequate if it instead fulfilled these four requirements for the concept of rightness, if it took this concept to be the more fundamental concept of moral thought.

our minds some idea—describable in other terms or not—of the property we are claiming that things in the world objectively have.

(2) An explanation of the way in which things in the world objectively satisfy our concept of goodness. A plausible realism must explain how a judgment-independent connection exists between our concept of goodness and the things it claims satisfy this concept. For such a connection to be judgment-independent, it cannot simply depend on the fact that we call certain things "good," or that we tend to associate them with our concept of goodness, even if we do so in a fairly systematic way. We must find something *within the concept of goodness itself* that makes it apply to certain things and not others. This could be a resemblance between our idea of goodness and the intrinsic properties of certain things in the world, as is suggested by a Lockean theory of the application of concepts, or it could be something else. Whatever it is, it must answer the following question: In what way does our representation of things as good reflect how these things are in themselves?

(3) An explanation of the way in which we can come to *know* which things objectively satisfy our concept of goodness. On the assumption that Criterion 2 is met—that there is a judgment-independent connection between our concept of goodness and something in the world—how is it that we have come to know this? It won't do for us to believe that certain objects or actions satisfy our concept of

goodness in some way epistemologically inaccessible to us, for though it's theoretically possible for them to do this, it would give us no justification for *believing* that they do. For us to be justified in believing that something in the world objectively satisfies our concept of goodness—i.e., for us to be justified in believing in realism—there must be some evidence available to us that points to this fact. What kind of evidence is this, and what faculties of ours enable us to recognize it?

(4) An explanation of why we are often mistaken about what is good. Given that we have some ability to recognize which things objectively satisfy our concept of goodness, how is it that we nevertheless frequently get things wrong, as evidenced by the existence of disagreement? Not only does an adequate realist theory need to explain why we are often mistaken about the moral facts in individual cases, but it also needs to meet Criteria 1 through 3 in such a way that we can see why, if it is the correct metaethical theory, it has not been obvious to everyone all along. Whatever account a realist theory ultimately gives of our concept of goodness and its connection to good things in the world, it is going to have to be compatible with our having been confused about the subject for a very long time.

Perhaps these seem rather obvious requirements for a realist metaethical theory. Yet intuitionism, minimal realism, synthetic naturalism, and ideal-observer theories with realist ambitions all fail on a majority of these points. A new approach is needed that takes very seriously all of these metaphysical and epistemological requirements.

The reader may have noticed that I haven't criticized analytic naturalism in this chapter. The reason for this is that I believe analytic naturalism—or, more generally, analytic descriptivism—actually holds the key to meeting the above criteria. Though previous versions of analytic descriptivism have not done this satisfactorily, I believe the potential exists. Granted, analytic descriptivism has long been out of favor, due primarily to the popularity of G. E. Moore's Open Question Argument. I will spend a major portion of Chapter 4 addressing the Open Question Argument; I will argue that it is not devastating, and in fact that the particular brand of analytic descriptivism I endorse has a very plausible way of neutralizing it. For now, however, I turn to explaining the basic premise of my view.

PART II

A ROBUST REALIST THEORY

AND ITS METAETHICAL DEFENSE

CHAPTER 3

NORMATIVE QUALIA

In an effort to demonstrate the superfluity of "objective" value, Hare asks us to conduct the following thought experiment:

> Think of one world into whose fabric values are objectively built; and think of another in which those values have been annihilated. And remember that in both worlds the people in them go on being concerned about the same things – there is no difference in the "subjective" concern which people have for things, only in their "objective" value. Now I ask, "What is the difference between the states of affairs in these two worlds?" Can any answer be given except "None whatever"?[89]

Hare asserts that, even if there were such things as objective values, their disappearance would make absolutely no observable difference, and we would go on caring about exactly the same things as before. In light of this, we seem to have no

[89] R. M. Hare, "Nothing Matters," in *Applications of Moral Philosophy* (London: Macmillan, 1972), 47.

reason to worry about "objective value," and we seem to have every reason to embrace antirealism, a moral philosophy which emphasizes the things that actually matter to us. What is lost with regard to "objectivity" seems to be more than made up for by the urgency and immediacy of our subjective concerns.

I believe, however, that Hare's dismissal of objective value in favor of subjective, felt concern is based on a false conception of objective values and subjective concerns as necessarily independent. Hare's antirealism is motivated by a repugnance toward the idea that there could be some hidden intrinsic value of things to which our lived human concerns have no connection. I agree with him that that sort of hidden objective value is superfluous with regard to both our practical and theoretical concerns. What Hare seems not to realize, however, is that objective value could be built into the fabric of the world in such a way that it *does* make a significant difference to people's felt concerns. To put it very simply, because people are part of the world, if value is built into the world's fabric, it could be built into those parts of the world that are people. That is, value, if it is part of the intrinsic nature of the world, is possibly part of the intrinsic nature of persons. We could be connected to the objective value of the world in virtue of being part of that world, and embodying some of its value in ourselves.

Consider again the criteria for a plausible realism given at the end of the last chapter. The thread connecting them all is the need for an explanation of the way in which our concept of goodness is related to objective value in the world. The criteria ask: What is this concept? How do things in the world objectively satisfy it? How do we come to know this? And why do we sometimes get it wrong? The problem that has stymied metaethicists is understanding how it's possible that anything objective and empirical could possibly satisfy our concept of goodness.[90] How could some concrete piece of the world directly embody to-be-promoted-ness? (And if it does, why has it taken us so long to notice?)

I believe the solution to this puzzle lies in the realization that not only are human beings objective parts of the universe, but so are their mental lives. If value is part of the fabric of the universe, it may not reveal itself to us through our eyes or ears, or through the results of elaborate physical experiments interpreted with the help of lengthy equations and supercomputers. Rather, value may reveal itself directly, through the nature of our mentality. It may be that value is less like the charge or spin of an electron and more like the quality of redness: not the redness that is constituted by certain reflective properties of a surface, but the redness that characterizes certain phenomenal experience. It may be that value is a phenomenal property, albeit a very special one, with particularly important ramifications.

If realism is to explain how we come to have a concept of value and how we apply it at all accurately, it must explain how value exists in such a way that it is judgment-independent and yet also closely related to the human mind. This can be

[90] The difficulty in imagining this motivates arguments like J. L. Mackie's argument from queerness. See his *Ethics: Inventing Right and Wrong* (London: Penguin, 1977), 38-42.

done if value is actually a phenomenal property: a quale. The nature of a “value” quale will not depend on our judgments about it, any more than the natures of color qualia depend on our accepting certain propositions about them. And yet at the same time there is a fairly straightforward story to be told about how, if value is a phenomenal quality of experience, we come to have knowledge of it.

Positing such a close connection between objective value and the human mind, in the form of phenomenal experience, is the foundation of my realist theory. I will spend this chapter explaining and defending the central premise of this approach: that there are indeed normative phenomenal qualities which embody not purely subjective, but objective, value. My first task will be to explain exactly what I have in mind in referring to these normative phenomenal qualities. My primary example of an experience which includes a normative phenomenal quality will be the experience of pain. I will consider the possibility that the “badness” of pain is reducible to its disposing us to avoidance behavior, but I will argue in the end that there is a phenomenal badness in addition to this behavioral disposition and that it is this phenomenal badness which is intrinsically normative.

I will go on to argue that this phenomenal badness is also instantiated in experiences besides that of pain: for instance, in the experience of disliking something, such as a food, and in the experience of emotional distress. And I will argue that there is a contrasting phenomenal goodness instantiated in all of our positive phenomenal experiences.

I will then turn to discussing whether the value that is present in phenomenal experience can really be considered objective: i.e., judgment-independent. To understand why the value of normative phenomenal experience is objective, it will be necessary to understand that normative phenomenal experience is fundamentally *non-intentional*, that is, that its normativity is not primarily directed towards another object but first and foremost characterizes the normative experience itself. I will thus argue for the fundamentally non-intentional nature of normative phenomenal experience and then close the chapter with a brief preview of the way in which the objective value of normative phenomenal experience will allow us to construct a realist theory meeting the criteria outlined in Chapter 2.

I. What are normative qualia?

Moral realism claims that normativity is a feature of the world independently of anyone's thoughts, beliefs, or attitudes about it. This does not mean, however, that a realist must claim that normativity is primarily a feature of actions or physical states of affairs. My view denies this, and this separates my view from most other realist theories. It even separates it from other theories which appeal to moral phenomenology, such as John McDowell's, which focuses on which intrinsic qualities of things merit certain phenomenal responses.[91]

[91] See John McDowell, "Values and Secondary Qualities," in Stephen Darwall, Allan Gibbard, and Peter Railton, eds., *Moral Discourse and Practice: Some Philosophical Approaches* (Oxford: Oxford

Rather than claiming that normative phenomenology reveals to us the goodness or badness of actions or of physical objects or states of affairs, I claim that it reveals to us first and foremost the intrinsic goodness and badness of the phenomenology itself. I claim that intrinsic goodness and badness are *phenomenal qualities of experience*, and that it is facts about these qualities of experience which constitute all the normative facts there are. On my view, the intrinsic goodness of a state of affairs depends solely on the amounts of positive and negative experience had by all of the conscious beings that are part of that state of affairs, and on how strongly positive and negative their experiences are. The goodness of an object, an action, or a disposition depends wholly on its conduciveness to bringing about intrinsically good states of affairs defined in this way.

I propose that objectively good and bad phenomenology is found within our common experiences of pleasure and pain, liking and disliking, happiness and depression, peacefulness and fear, and meaningfulness and despair. I propose that in all of our everyday experiences of pleasure, of liking something, of being happy, of being peaceful, and of feeling that life is meaningful, there is a common positive feeling: a feeling of goodness *tout court*. What we are experiencing, I propose, is the most basic, elemental type of positive value. We experience goodness directly and immediately as one of the defining properties of these feelings. A feeling that does not

University Press, 1997), 201-13. For a theory which appeals to moral phenomenology in a similar way, see Maurice Mandelbaum, *The Phenomenology of Moral Experience* (Glencoe, Ill.: Free Press, 1955).

feel *good* cannot qualify as pleasure. Nor can it qualify as happiness. Peacefulness without its positive aspect is better described as boredom. Defining pleasure, happiness, and peacefulness requires referring to some positive quality present in the experiences. And it is this positive phenomenal quality that I propose is identical with the property of intrinsic goodness.

On the other hand, in experiences of pain, of disliking something, of being depressed, of being afraid, and of feeling hopeless or despairing, I believe we feel something that is the direct opposite of this quality. In all of these latter feelings, there is a common element of negativity. I submit that, when we have these negative feelings, we directly experience intrinsic badness.

It is extremely important that it be understood that I am not suggesting that our normative phenomenology represents some further realm of normativity, that it somehow acquaints us with normative properties that also exist detached from phenomenal experience, perhaps in actions or in non-mental states of affairs. Note that I have not given as paradigm examples of normative phenomenology the experiences of disapprobation, indignation, or shame, which are sometimes understood to be feelings which signal the objective badness of the actions they are said to be “about.” While I do believe that experiences of disapprobation, indignation, and shame contain an instantiation of the negative phenomenal quality with which I am concerned, I want to make clear that this negative phenomenal quality alone is much more basic than the complex experiences of disapproval, indignation, and shame and does not necessarily represent anything else objectively bad. I will eventually, in Chapter 7, discuss the more complicated cases in which normative phenomenal experiences serve as indicators of the instrumental value of actions or objects with which they are associated, but note here that these are not the most basic cases of normative phenomenology on my view, and that even in these cases, the normative experiences are themselves good or bad, and it is their own intrinsic value or disvalue that allows them to represent the instrumental value or disvalue of other things.

Normative phenomenal experience, I believe, does not *necessarily* point us toward anything else normative. Its intrinsic normative nature is not essentially representative of any mind-independent normative reality (though its normative nature *is judgment-independent*: recall the distinction between mind-independence and judgment-independence drawn in Chapter 1). Rather, the phenomenal experience itself is good or bad, and any representative function it may play is purely secondary and contingent. (I will argue for the fundamental non-intentionality of normative phenomenology in Section V of this chapter.)

My proposal is that *intrinsic goodness and badness just are felt qualities*. One of the things I normally feel when I am in pain is an intense unpleasantness. This unpleasantness, I propose, is intrinsic badness itself. It is bad to feel that way (all else being equal). It is bad to have experience with that negative quality, simply in virtue of the nature of the feeling itself. *The feeling itself is badness*. When I feel pleasure, on the other hand, I feel goodness itself. It is good to feel that way, simply in virtue of the character of the feeling. *The feeling is goodness*.

It might be objected that I am confounding two separate senses of the terms 'good' and 'bad', that I am mistaking *feeling* good and bad for *moral* goodness and badness. Let me say that I am well aware of the many senses that can be given to the terms 'good' and 'bad'—I will outline several of them in the next chapter—and that my identification of intrinsic moral goodness with felt goodness is completely intentional. I don't think it's an accident that we use the term 'good' to refer to both phenomenal and moral properties. I believe experiencing felt goodness provides us with the basic qualitative content of our concept of intrinsic moral goodness. Felt goodness *is* moral goodness of the most basic kind; it is the basic objective value that gives meaning to moral discussion and action. And this conceptual relation between phenomenal goodness and moral goodness is the key to a robust moral realism, to a realism which locates goodness within the empirical realm.

I have listed some examples of experiences in which normative phenomenal experience is frequently or even always present: experiences such as pleasure and pain, happiness and depression. But some of these terms are more ambiguous than others, and not every experience to which we would apply one of these labels contains a normative phenomenal component. Specifically, not everything we call "pain" includes an instantiation of the phenomenal quality of badness. Directing the reader's attention to just that element of phenomenal experience that is the normative feeling may be somewhat difficult, in part because our language doesn't currently have a term for it. Our vocabulary for describing any phenomenology is quite limited, and, with the terms presently in use—terms like 'pleasure' and 'pain'—the normative aspect of experience gets lumped together with other phenomenal qualities that often accompany it, but which are non-normative. For this reason, I'm going to introduce the term 'normative qualia' to refer to just those phenomenal qualities whose feel is normative, whether positive or negative. But, given this term, I have to make clear exactly which aspects of phenomenal experience I mean to be picking out by using it.

Unfortunately, normative phenomenology can't be accurately described by analogy with anything else. This is the situation with all basic phenomenal qualities. Just as there's no way to convey to a congenitally blind person what experiencing color is like, or to convey to the congenitally deaf the nature of the tones, timbres, and harmonies produced by an orchestra, neither could the nature of normative phenomenal qualities be conveyed to someone who had never experienced them. I think I am safe in assuming, however, that all of my readers will have experienced normative qualia.[92] And so while I can't proceed by analogy, I can describe more

[92] Perhaps there are persons who do not experience normative qualia, but I am inclined to think that such qualia are fairly integral to the normal functioning of the brain, and that someone who did not have the capacity to experience them would have a drastically different mentality, manifested in drastically different behavior. Antonio Damasio's research on persons who have great difficulty making decisions has led him to hypothesize that their difficulty is due to an impairment of their emotions. (See Antonio Damasio, *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain* [New York: Harcourt Brace & Company, 2003], Ch. 4.) I imagine the situation would be even worse were the patients unable to feel any sort of normative qualia at all. A claim like this relies on phenomenology playing a causal role, but, as yet, I see no reason to assume that it does not.

specifically the various situations in which I believe most people experience normative phenomenology, in hopes that the reader will eventually recognize a quale common to their experiences on all these occasions.

II. The experience of pain

The logical starting place is the experience of pain. Pain is the most frequently cited example of something that is objectively, intrinsically bad. The badness of torture (especially torture just for the fun of it) is probably the moral value the most easily agreed upon and the most frequently appealed to in arguments for the self-evidence of certain moral truths.[93] There is something about the experience of pain which convinces many people of moral realism. I believe this thing is a normative quale, an intrinsically negative phenomenal quality whose instantiation is normally (though not always) a component of the experience of pain. I believe that part of the usual phenomenology of pain is an instantiation of the phenomenal quality of undesirability—of badness—and that instantiation of this quality is bad no matter what judgments anyone makes about it. Experience of this quality is what I believe leads many people to assert with such confidence that pain is objectively bad.

Of course, the experience of pain does not convince everyone of moral realism. Antirealists, despite having what we can only assume is a qualitatively similar

[93] See, for example, Nagel, *The View from Nowhere*, 156-62; Shafer-Landau, *Moral Realism*, 248; and Nicholas Sturgeon, "Moral Explanations," in Geoffrey Sayre-McCord, ed., *Essays on Moral Realism* (Ithaca, NY: Cornell University Press, 1988), 229-55.

phenomenal experience of pain, do not conclude that any part of pain is objectively, intrinsically bad, but say we are simply wired in such a way as to *take* experiences like pain to be bad. Sharon Street, for instance, proposes that "[p]ain is a sensation such that the creature having the sensation unreflectively takes that sensation to count in favor of doing whatever would avoid, lessen, or stop it,"[94] but, she emphasizes, "the badness of pain does in fact depend on our evaluative attitudes."[95] Christine Korsgaard expresses a similar opinion, writing that "someone who says he is in pain is not describing a condition that gives him a reason to change his condition. He is announcing that he has a *very* strong impulse to change his condition." She goes on to say that "[t]he painfulness of pain consists in the fact that these are sensations that we are inclined to fight. … Pain is not the condition that is a reason to change your condition…. It is our *perception* that we have a reason to change our condition. Pain itself is not a reason at all."[96]

Antirealists like Street and Korsgaard say that, while we are wired to dislike pain sensations, and to conscientiously avoid them, both for ourselves and for those with whom we sympathize, this does not mean there is anything intrinsically bad about the sensations of pain themselves. Korsgaard insists,

> Pain really is less horrible if you can curb your inclination to fight it. This is why it helps, in dealing with pain, to take a tranquilizer or to lie

[94] Street, "A Darwinian Dilemma for Realist Theories of Value," 146.

[95] Ibid., 145.

[96] Christine M. Korsgaard, "The Sources of Normativity," in *The Tanner Lectures on Human Values*, vol. 15, edited by Grethe B. Peterson (Salt Lake City: University of Utah Press, 1994). Excerpted in Darwall, Gibbard, and Railton, 389-406, p. 402.

> down. Ask yourself how, if the painfulness of pain rested just in the character of the sensations, it could help to lie down? The sensations do not change. Pain wouldn't hurt if you could just relax and enjoy it.[97]

According to antirealists, pain sensations trigger a certain avoidance reaction, but they have no intrinsically negative phenomenal character.

Given that not everyone is agreed that the badness of pain is intrinsic to its phenomenal character, I think it will be helpful to spend some time carefully reflecting on the phenomenology of pain. I believe that in doing so we may be able to understand better the reason antirealists like Street and Korsgaard deny pain's intrinsic badness, and yet come to see this denial as mistaken. What I hope to show is that, despite the fact that there are some elements of the experience of pain which are not bad in themselves and which we only take to be bad because we react to them in a certain way—here the antirealist is correct—there is nevertheless another element of the experience of pain which *is* bad in itself, whose badness does not depend on any further reaction of ours which *takes* it to be bad.

The first thing to understand is that pain is normally a compound experience made up of at least two components. On the one hand, there are the sensations of nociception. These are the sensations that vary with the kind of harm done to one's body. They will be different if one is stabbed, burned by something hot, or burned by a chemical. They may be sharp or dull, pulsing or constant, and felt to be at different

[97] Ibid.

locations in the body. In addition to these sensations of nociception, however, the experience of pain also normally includes a feeling of *badness*.

This division of the experience of pain into at least two basic parts is widely accepted by scientists because of the existence of cases in which the sensations of nociception are separated from the feeling of badness. This separation occurs, for instance, under the influence of certain painkillers: namely, opiates such as oxycodone and morphine. Users of these drugs relate that they don't make the pain go away so much as they make one no longer care about it. This phenomenon—called "reactive disassociation"—is well-documented and is also known to occur as a result of prefrontal lobotomies and leucotomies, and as a result of lesions on the anterior cingulate cortex.[98]

The existence of two distinguishable, and sometimes separable, components of the experience of pain is sometimes used by philosophers as evidence that pain sensations are not bad independently of our reactions to (or judgments about) them. Richard Hall, for instance, cites reactive disassociation as a demonstration that one's negative reaction to pain can be removed without disturbing the phenomenology, thus proving that pain is not intrinsically bad.[99]

[98] See, for instance, D. D. Price, A. Von Der Gruen, J. Miller, A. Rafii, and C. A. Price, "Psychophysical Analysis of Morphine Analgesia," *Pain* 22 (1985): 261-69; E. L. Foltz and L. E. White, "Pain 'Relief' by Frontal Cingulotomy," *Journal of Neurosurgery* 19 (1962): 89-100; J. M. Gybels and W. H. Sweet, *Neurosurgical Treatment of Persistent Pain* (Basel: Karger, 1989); and P. D. Wall, *Pain: The Science of Suffering* (New York: Columbia University Press, 2000).

[99] Richard J. Hall, "Are Pains Necessarily Unpleasant?", *Philosophy and Phenomenological Research* 49, No. 4 (June 1989): 643-59, p. 651.

One of the most recent arguments to the effect that the badness of pain sensations depends on our normative judgments is given by Street.[100] She lays out a dilemma for realists about the badness of pain: either they can say that having a negative evaluative reaction to the sensations of pain is necessary to these sensations' being pains, or they can say that it is not so necessary. If the realist chooses the second tack, there is the conceptual possibility of some individual's having a *positive* reaction to pain sensations, of his coming "positively to *enjoy* the sensation in question."[101] But the realist is forced to say that, in such a case, the pain sensations are nevertheless intrinsically bad, a highly counterintuitive result. A realist who wants to avoid saying that pain sensations which are enjoyed are still bad is forced to embrace a view on which a negative reaction to pain sensations is essential to their being pains: if one has a positive reaction to them, then they don't qualify as pains. But this, Street says, is to accept that the badness of pain sensations depends on our evaluative attitude towards them, and thus to accept antirealism about the badness of pain. She writes,

> In order to salvage his or her view of pain as bad independently of our evaluative attitudes, the realist must admit that pain's badness depends on its being a sensation such that the creature who has it is unreflectively inclined to *take* it to be bad. But this, in turn, is just to admit that its badness depends in an important sense on our evaluative attitudes—in particular, on our being unreflectively inclined to take it to be bad. Pain may well be bad, in other words, but if it is so, its badness hinges crucially on our unreflective evaluative attitudes toward the sensation which pain is.[102]

[100] Street, "A Darwinian Dilemma for Realist Theories of Value," 144-52.
[101] Ibid., 149.
[102] Ibid., 151.

Such arguments for antirealism about the badness of pain rely on a fairly simple, but faulty, strategy. They note that there is a part of the experience of pain which includes no feeling or judgment of badness, and another part which is this feeling or judgment of badness. They note that obviously the part that includes no feeling or judgment of badness is not intrinsically bad, and then conclude that the badness of pain as a whole is not intrinsic, because it depends on the other part of the experience, which is our reaction to the first. What they don't consider is that this other part of the pain experience might be the experience of an additional phenomenal quality and that this quality might itself be intrinsically, judgment-independently bad. Sometimes—e.g., in cases of reactive disassociation—no such quality is present, but, in those cases where we do react negatively to sensations of nociception, I propose that this is explained by our *experiencing a negative normative quale* along with these sensations.

That is, pain is normally bad because it is normally not just the sensations of nociception; it is a composite of these sensations and an instantiation of the phenomenal quality of badness. (In some cases, the combination of nociceptive sensations and the phenomenal quality of badness may also be accompanied by an experience of the phenomenal quality of goodness. Masochists, for instance, arguably experience pleasure—or at least relief from psychological suffering—when they experience pain, or when they anticipate its relief.) But although a pain's being bad requires its including an instantiation of this additional negative normative quale, the badness of experiencing this additional negative normative quale is in the feel of the quale itself and not dependent on any reaction we have to it or any judgment we make about it. For this reason, the badness of experiencing this quale provides an objective, judgment-independent reason for experience of it to be avoided, lessened, or eliminated.

III. Phenomenology or behavioral disposition?

But is there reason to think that the negative reaction eliminated by opiate analgesics includes an intrinsically bad phenomenal experience, that it isn't just a behavioral disposition to avoidance, as the antirealists would have it? Austen Clark affirms that pain as normally talked about has at least two parts—its sensory character and its undesirability—but he maintains that its undesirability is exclusively a motivational/dispositional feature and has no phenomenal aspect. He discusses this in an essay titled "Painfulness is Not a Quale."[103] Yet while Clark makes a good case for aversion's being an essential component of pain, he does not present a compelling case against there being something it's like to *feel* aversion. He claims that the aversiveness of pain is necessarily a relational property, a relation in which a sensation

[103] Austen Clark, "Painfulness is Not a Quale," in Murat Aydede, ed., *Pain: New Essays on Its Nature and the Methodology of Its Study* (Cambridge, MA: MIT Press, 2005), 177-97.

stands to one’s motivational states, and he says that the fact that it is relational means that it cannot be a quale. He writes that

> one could have two instances of mental states that are qualitatively identical, that share all the same sensory qualia, yet which are not equally aversive. Surround that same sensory state with a different constellation of preferences, and this second instance of the same state may not be equally painful. So painfulness is not a quale. It is at best a motivational disposition occasioned by a quale. To paraphrase Wittgenstein and Anscombe (Anscombe 1957, p. 77): no immediate phenomenological quality could be an aversion, because it cannot have the consequences of aversion.

What Clark doesn’t even consider is that there might be a way that it *feels* to have an aversion: that one’s dispositions might have a phenomenology, or that one’s phenomenology of aversion might be a necessary cause of one’s dispositions.

In his attempt to show that painfulness is not a quale, Clark offers the following criteria for something’s being a quale: (1) it must be instantiated in various sensory episodes, and (2) that in virtue of which two sensory episodes instantiate the same particular quale cannot be defined in any functional or behavioral terms. These criteria seem acceptable to me, but Clark’s method of argument from them does not work. He argues that the aversion that is an essential part of pain is purely functionally definable, and thus that painfulness is not a quale. But there are two possibilities he doesn’t consider. First, aversion may be purely functionally definable and essential to being a pain, and yet it may not be *sufficient* for something’s being a pain; there may be an additional quale necessary—a quale of badness. Second, the type of aversion essential to being a pain may *not* be functionally definable. It might be that aversion *behavior* isn't what's necessary for one's having a pain. Perhaps what's necessary is an aversion that is *felt*: i.e., a phenomenal manifestation of aversion. I incline to this second view.

Unfortunately, there is no straightforward way to present evidence for a phenomenal manifestation of aversion, since there's no way literally to point at the instantiation of a phenomenal quality as proof of its existence. But there are some thought experiments (as well as scientific experiments) which may help to make the existence of normative qualia clearer.

Consider, for instance, the case of someone who is stuck with a pin and responds by smiling, sighing, and asking to be pricked again. Now imagine that, despite these reactions that normally indicate pleasure, our subject is actually in quite a bit of pain each time she is pricked. In fact, imagine that she feels an intense desire that the pricking stop, and yet, for some reason, her outward behavior occurs completely disassociated from this desire. No matter how intensely bad the pricking feels, she can't manage to control her own behavior and line it up with her felt desire that the pricking stop. The fact that we can imagine what it would be like to be someone trapped in a body with dispositions contrary to one's felt desires seems some evidence for there being both a behavioral aspect and a phenomenal aspect to our aversions.

Furthermore, although such an extreme case of disassociation between one's felt desires and one's behavioral dispositions may not be physically possible, cases of lesser degrees of disassociation have actually been observed. C. W. Sem-Jacobsen reports that, when he was conducting experiments in which he stimulated subjects' brains with implanted electrodes, he encountered a subject who smiled, laughed, and generally seemed to enjoy being stimulated at a brain site thought to be a "strong, positive 'pleasure region,'" but who one day suddenly became angry and said she was "fed up," and "did not enjoy these stimulations at all."[104] This subject's behavioral dispositions seem to have been at odds with her feelings toward the stimulations, making it seem less likely that these two things are equivalent.

Consider, too, cases in which there is a disposition to avoidance behavior and yet no associated phenomenal experience, either negative or positive. It seems quite easy to imagine someone who systematically avoids certain things and yet, when forced to encounter them, *feels* no particular aversion to them. We can also imagine a situation in which someone shows every sign of being in pain when a certain event occurs and yet doesn't actually feel any aversion to the event. We can imagine that their pain behaviors are simply a matter of reflex and don't reflect any conscious discomfort.

[104] C. W. Sem-Jacobsen, *Depth-Electrographic Stimulation of the Human Brain and Behavior* (Springfield, Ill.: Charles C. Thomas, 1968), 131-2.

Actual cases of these sorts of reflexive responses have been documented by M. M. Morgan, M. M. Heinricher, and H. L. Fields, among others.[105] Robert Coghill summarizes their findings thus:

> [N]ociceptive responses in the form of reflex withdrawals can be elicited when a noxious stimulus is applied to the…lower limb of a human…. These withdrawal responses can also be exquisitely modulated by stimulation of other body regions to easily produce the outward appearance of a logical, conscious decision about the best motor plan for escaping from the noxious stimulus (Morgan, Heinricher, and Fields 1994). Obviously, in the case of the spinal cord transected human subject, such reflex withdrawals occur with no verbal report of a pain experience, indicating that the nociceptive information has not been sufficiently processed to elicit a subjectively available conscious experience.[106]

Since in these cases there is avoidance behavior without conscious aversion, the two cannot be equivalent.

This does not yet prove, however, that conscious aversion must have a phenomenal aspect. Perhaps conscious aversion is only a disposition—but a mental, brain-based one, rather than one directed purely by the spinal cord. It might be suggested that, even in a case where one's outward behavior implies that one is experiencing pleasure, one could inwardly have a contrary disposition, a disposition manifested in one's *thoughts*. For example, one might be contemplating various ways in which to avoid a further needle-pricking while at the same time none of these

[105] M. M. Morgan, M. M. Heinricher, and H. L. Fields, "Inhibition and Facilitation of Different Nocifensor Reflexes by Spatially Remote Noxious Stimuli," *Journal of Neuropsychology* 72 (1994): 1152-60.

[106] Robert Coghill, "Pain: Making the Private Experience Public," in Murat Aydede, ed., *Pain: New Essays on Its Nature and the Methodology of Its Study* (Cambridge, MA: MIT Press, 2005), 301.

reflections leads to outward behavior because of an abnormality in the nervous system. What I've been calling the "phenomenology" of badness could be explained away as a disposition to reflect consciously on avoidance strategies, to *plan mentally* for avoidance.

But consider now a case more extreme than those we've discussed so far: that of someone who is paralyzed and who is also severely mentally handicapped, to the point that he doesn't even understand the concept of cause and effect, and so can't understand that some action of his could affect his future experiences. It seems that even someone like this could be acquainted with the badness of pain. He would not be capable of contemplating escape strategies (much less of carrying them out), and perhaps he would not even be capable of conceptualizing the fact that escape is what he needs. Nevertheless, it seems he could experience the badness of pain.

In fact, there do exist thousands of human beings in the world unable to understand cause and effect—newborn babies—and yet despite the fact that newborns can't plan to avoid their pain, and for awhile cry only from reflex, we tend to believe that they can nevertheless *feel* pain and discomfort. We also tend to believe that many species of animals whose brains are not well enough developed to engage in planning are able to experience pain. This is some evidence that we are acquainted with a phenomenal aspect to badness which exists in addition to whatever avoidance dispositions—physical or mental—one may manifest.

It is not conclusive evidence, however. It could be suggested that, when we are imagining being in "phenomenal" pain, we aren't doing anything as cognitive as making plans, but neither are we bringing to mind any uniquely normative phenomenology. It might be suggested that we are imagining being in a very subtly different bodily state, one that includes grimacing, tensing muscles, clenching teeth, etc. This was the hypothesis of William James.[107] Perhaps, while we may not realize that it is a bodily state that we are conjuring up when we imagine being in pain, if we pay attention to what is actually happening in our bodies when we try to imagine "normative phenomenology," we will see that we are tensing our muscles, etc. We will see that we *are* in fact imagining a bodily state, whether we know it or not. Perhaps this bodily state is what being in pain is—nothing phenomenologically unique, and certainly nothing uniquely normative.

This is a difficult objection to which to reply, because I do believe that our phenomenology is very closely tied to our bodily states. The research of neuropsychologist Antonio Damasio indicates some very tight connections between bodily states and emotions, for example.[108] And yet I believe that, despite their tight causal connection, the negative phenomenology of pain is distinct from its bodily manifestations. Some evidence for this is found in the case described by Sem-

[107] William James, *What Is an Emotion?* (1884).

[108] For non-technical accounts of Damasio's work, see his *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (New York: Harcourt Brace & Company, 1999) and *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain* (New York: Harcourt, Inc., 2003).

Jacobsen, cited above. Also, Ian Glynn, in *An Anatomy of Thought: The Origin and Machinery of the Mind*, mentions evidence which appears to contradict James's theory by showing that "the same bodily effects may occur in very different emotional states, and even in non-emotional states so that something more is needed to account for particular emotions." Glynn cites a study by Stanley Schachter and Jerome Singer in which participants were injected with adrenaline but it was discovered that, in order for the adrenaline to affect their mood, they also had to be placed in a situation that was at least "slightly cheerful or slightly irritating." That is, a positive or negative change in mood was not produced by the adrenaline alone. (Nor was it produced by the situation alone, as shown by the lack of reaction in those injected with an inert substance.) A situation with a positive or negative valence had to be added to the injection of adrenaline in order to determine the nature of the emotion experienced.[109]

I believe that our ultimate attitude toward our bodily state is a function of whether instantiations of positive or negative normative qualia accompany it. Without this normative phenomenology, there might still be the physical manifestations of pain, and even the conscious awareness of these physical manifestations, but they would not feel *bad*. *Felt* badness seems to exist *in addition* to all of our physical manifestations of pain.

[109] Ian Glynn, *An Anatomy of Thought: The Origin and Machinery of the Mind* (New York: Oxford University Press, 1999), 335. Glynn cites Stanley Schachter and Jerome Singer, "Cognitive, Social, and Physiological Determinants of Emotional State," *Psychological Review* 69 (1962): 379-99.

But there is yet another argument for the existence of positive and negative qualia, one that doesn't depend on introspection or on people's reports of their feelings. It seems that we ought to *expect* our dispositions of attraction and aversion to be manifested in phenomenal experience, for the simple reason that so much other important information that our brains use in conscious decision-making is manifested in this way. Granted, scientists and philosophers of mind remain puzzled about *why* our decision-making is accompanied by phenomenal experience, but it is relatively well accepted that phenomenal experience is present when and where an ability consciously to manipulate information is present. Conventional blindsight patients, for example, who have lost their visual phenomenology, are not able consciously to carry out tasks requiring visual information about their environment, though they do have the remarkable ability to carry out some of these tasks *unconsciously*. Thus they may be able to navigate around the furniture in a room, without being able to describe the furniture. Interestingly, some blindsight patients retain some form of phenomenal experience, but of a surprising nature. In a case reported by Larry Weiskrantz and Elizabeth Warrington, a blindsight patient stated that, despite lacking all visual sensations, he was able to discriminate between a circle and a cross due to sensations of "smoothness" and "jaggedness."[110]

[110] L. Weiskrantz, *Consciousness Lost and Found* (Oxford: Oxford University Press, 1997); and L. Weiskrantz, E. K. Warrington, M. D. Sanders, and J. Marshall, "Visual capacity in the hemianopic field following a restricted occipital ablation," *Brain* 97 (1974): 709-28.

I would argue that if our brain appears dependent on (or at least very fond of) phenomenal experience when it comes to conscious manipulation of information—and if this extends even as far as producing qualia of smoothness and jaggedness—then we would seem to have good reason to expect something as important to decision-making as attraction and aversion to have a phenomenal presence as well. Attraction and aversion are what tell us how to react to all of the other information we have about our environment. Our visual system may let us know that a lion is up ahead, but unless we also have access to our brain's verdict about whether the proximity of a lion is good or bad, this purely visual information is useless to promote our survival. Our brains are able to make judgments about whether a particular situation is likely good or bad for us (based on instinct, past experience, or learning from others), but in order for these judgments to play a part in our conscious decision-making—in order for us to be able to reflect on our aversion to the lion and consciously contemplate the best escape strategy—it would seem consistent with our other knowledge about conscious thought that we require a phenomenal manifestation of our aversion: that we need a phenomenal marker that conveys to us the brain's verdict of "Bad!". It doesn't seem sufficient for our brain automatically to put us in certain bodily states or produce certain actions, since it is to our advantage in most cases to reflect consciously on just *which* actions are called for. But in order to reflect in this way, we have to be consciously aware of a very general message from our brain saying "Bad!" or "Good!" so that we know which direction our reflection should take. Based on our current knowledge of the way consciousness works, it seems likely that in order to be available to conscious decision-making in the widest possible way, this message needs to be phenomenal.

It is a further question, of course, exactly what quality this phenomenology has, if it exists. Must it have a quality that is objectively, intrinsically bad? Or might the phenomenology of aversion just have some arbitrary quality that happens to have become employed for the purpose of making our aversive dispositions available for conscious reflection? This is a hugely complicated question—one that I can't get into in depth here—but I do want to say that, while it might not matter which phenomenal quality represents to our consciousness a certain ratio between different wavelengths of light, so that it might seem that there's nothing wrong with its being completely arbitrary that apples produce in us the color phenomenology they actually do rather than the color phenomenology bananas do, it could be that the phenomenal quality informing us of our brain's verdict about attractiveness or aversiveness plays a more important role.

Granted, we know very little about the role phenomenology in general plays in our conscious decision-making, but most scientists and philosophers—epiphenomenalists being the exception—seem unsatisfied with the idea that it plays no causal role at all. I would suggest that, if we are interested in discovering a connection between phenomenal qualities and causal roles, we begin our hypothesizing with the phenomenal qualities of goodness and badness. These two qualities of phenomenal experience, more than any others, seem to have crucially different effects on our conscious decision-making. It seems that we could exchange the phenomenal qualities of red and yellow in our minds without affecting their usefulness in discriminating among objects in our visual field. We could, for example, still recognize bananas; it would simply be red qualia that we would associate with bananas rather than yellow. It doesn't seem, however, that if we associated a positive feeling with the things that our brains want us to avoid, and associated a negative feeling with the things our brains want us to seek out, that everything would necessarily go on working as before. The actual phenomenal qualities of feeling good and feeling bad seem as though they might have a causal influence on our conscious decision-making. I presently have no detailed hypothesis to offer to make this claim more scientifically compelling, just the intuition that if the intrinsic nature of a phenomenal quality were ever important, it seems it would be so in this case. Thus I think we shouldn't rule out the possibility that the phenomenal markers of attraction and aversion have importantly unique phenomenal qualities, ones that could in fact have ethical implications.

And indeed we ordinarily accept that this phenomenology has ethical implications. Not only do we affirm the badness of causing another person pain and the goodness of bringing another person pleasure, *ceteris paribus*, but the crucial question many people ask in deciding how to treat an organism of another species is whether it *feels pain*. Pain *behavior,* or avoidance behavior in general, isn't sufficient to cause us to take another creature's welfare into consideration if we have been convinced that they do not have the *feeling* of pain. It seems to me that a good explanation for our taking this criterion to be important is that there is something more to our phenomenal experience of pain than mere bodily sensations and aversive behavior. Our question about animals seems to presuppose the existence of intrinsically bad phenomenal qualities.

In this section, I have hoped to bring out more clearly what exactly normative qualia are and to provide intuitive and scientific evidence for their being instantiated in the (everyday) experience of pain. It is not easy to isolate the normative component of pain from all of its other sensory and dispositional elements, but the mere fact that our bodily states, our dispositions both physical and mental, and our phenomenal states are all closely intertwined does not mean that the normative part of the experience of pain must be reducible to any of the others.

IV. Normative qualia in contexts other than pain

To further the discussion as to whether there are intrinsically normative qualities of phenomenal experience, let me now turn to some examples of normative qualia instantiated in experiences besides pain. I mentioned earlier that normative qualia can be found in experiences of liking or disliking something. Consider the experience of eating a food that you find disgusting. I maintain that the experience contains an instantiation of a negative normative quale, located in the feeling of disgust itself. Instantiation of this quale is intrinsically bad. Its feeling is such that we would be better off, *ceteris paribus*, if we weren't feeling it. Judged only on its own intrinsic, qualitative merits, this is a feeling that ought not to be occurring.

What is interesting, and not often noted, is that the part of the phenomenology of tasting something disgusting that is *intrinsically* disgusting is not the taste of the food: its sweetness or saltiness, its texture or its smell. What is intrinsically disgusting is the accompanying negative phenomenology that it causes. As in the case of pain, where opiate analgesics can remove the normative quale while leaving behind many other sensations, the feeling of disgust can be removed while leaving intact the sensations of sweetness, saltiness, texture, and smell. Consider, for example, that the first time we try a new food or taste combination, we often dislike it, but, upon trying it again at some later time, discover we have come to like it. Through habituation, it seems that the same taste sensations can come to produce positive phenomenology rather than negative.

What I want to defend now is the thesis that the phenomenal quality that makes the experience of tasting a food disgusting is the same quality that makes pain feel bad. The proposal is that all of our negative experiences, whatever differences they may have, all have at least one qualitative aspect in common: their negativity. Negativity, I am suggesting, is itself a phenomenal quality—as is positivity—and it can come to be produced in conjunction with a practically infinite variety of other, non-normative phenomenal qualities, though the only thing intrinsically good or bad is the instantiation of the positive or negative quale itself.

Some philosophers have written what sound like objections to the claim that all negative experiences share some single phenomenal quality describable as "negativity" or "badness." For instance, Don Gustafson writes that "[i]t seems plain that humans categorize experiences as pains along several dimensions, with loads of differences as to sensory and affective dimensions. Pains seem too diverse to admit of a single defining or essential quale."[111] Granted, the phenomenology of pain appears to vary widely according to which type of nociceptors have been stimulated (those sensitive to mechanical, thermal, or chemical stimuli), at which location in the body, and with what intensity. However, the great variety of *other* phenomenal qualities that make up many of our pain sensations should not lead us to believe that there cannot be at least one phenomenal quality that they all share: the quality of badness.

Indeed, Gustafson acknowledges that we might pick out "hurtfulness" as a quality shared by all pains, but he doesn't pursue this tack because he says that, "[i]f we reduce [hurtfulness] to a single parameter, as we might in theory, it will pretty clearly cover more than pain. It will encompass any condition the agent or animal takes to be untoward, counter to its good, a threat to its projects or ongoing activity, or a state to which its host needs to attend."[112] Gustafson is interested in categorizing pain to the exclusion of all these other states. Thus, when he says "[p]ains seem too diverse to admit of a single defining or essential quale," he isn't ruling out the

[111] Don Gustafson, "Categorizing Pain," in Murat Aydede, ed., *Pain: New Essays on Its Nature and the Methodology of Its Study* (Cambridge, MA: MIT Press, 2005), 224.
[112] Ibid.

possibility that there is some quale that they all share. Rather, he thinks that whatever quality they all share—perhaps a common quality of hurtfulness—it is also a quality had by some experiences that we would not categorize specifically as "pains." But this is exactly my view: that pains share their quality of badness with many other experiences.

Hall also makes statements about the diversity of pain sensations that might seem contrary to my thesis, but when taken in context, they too are compatible with what I have been claiming. In an article titled "Are Pains Necessarily Unpleasant?", Hall answers this question in the negative. He writes,

> The natural view is that there is a common sensational quality, a common phenomenological feel, that unites all these pain sensations and accounts for our calling them all pains…. But what could that common phenomenal feel be? Think of how different the initial stabbing pain of a pin prick is from the dull ache of a bruise, and both of those from the feel of a burn or a cut. What phenomenal feel is common to all those pains? The obvious answer may seem to be: unpleasantness (or awfulness or horribleness). "Pains…have an intrinsic qualitative nature (a horrible one) that is revealed in introspection…." I disagree with this view. While I admit that pains are unpleasant, I hold that unpleasantness is not a phenomenal quality of pains. What does the unpleasantness of pain sensations consist in, then?
>
> The unpleasantness of pain sensations consists in their being disliked. The dislike of a pain sensation is a separate mental state, separate, that is, from the sensation.[113]

But while Hall says, "I hold that unpleasantness is not a phenomenal quality of pains," it is not clear whether he thinks this primarily because he thinks unpleasantness is not a phenomenal quality at all, or because he doesn't think it ought to be categorized as a

[113] Hall, 646.

quality specifically of pains or regarded as intrinsic to the physical sensations of pain. His primary justification for his view seems to be the empirical evidence he presents that “you could have exactly the same kinds of sensation as you have when you are cut, burned, or bruised, and they not be unpleasant.”[114] But this just begs the question against unpleasantness’ being a quality of the sensations one has in these situations. Merely showing that unpleasantness and sensations of nociception can come apart goes no way towards proving that unpleasantness is not also a phenomenal quality. Indeed, Hall tries to explain why we find pain sensations unpleasant by saying, “The distress we feel at perceiving our bodies getting damaged gets associated with the sensations which accompany those perceptions.”[115] He presumably understands “the distress we feel” as an attitude or judgment about the sensations accompanying bodily damage, but he gives no reason to think that this distress is not literally *felt*, that there is not an additional experiential quality that often accompanies sensations of nociception and which is intrinsically, objectively bad.

Apart from dismissing arguments from the diversity of nociceptive sensations and their only contingent causal relation to unpleasantness, we can also produce some positive evidence for the existence of a single phenomenal quality instantiated not only in normal experiences of pain but also in other experiences we would call “bad.” In an essay entitled “Social Pain, Support, and Empathy,” Jaak Panksepp summarizes

[114] Ibid., 643.
[115] Ibid., 647.

evidence for the claim that "emotional pain, such as that which accompanies grief and intense loneliness, does share some of the same neural pathways that generate the affective sting of pain."[116] He notes some obvious similarities, such as that both losing someone we love and being in intense pain cause us to cry. But what could seem merely superficial similarities turn out to run much deeper. For instance, the periaqueductal gray of the brain stem is known to help control the experience of physical pain, but it has also been shown to be associated with intense sadness,[117] and it is the region of the brain in which low levels of stimulation most easily evoke "emotional distress."[118] Research also points to another pain center—the anterior cingulate cortex—as involved in producing "distressing social feelings." Panksepp writes that "this brain area helps mediate social processes such as maternal behavior, social bonding, and separation calls (MacLean 1990; Panksepp 1998), which helps make sense of why this brain area may participate in the distress that arises from being socially ostracized (Eisenberger and Lieberman 2004; Williams 2001)."[119]

[116] Jaak Panksepp, "Social Pain, Support, and Empathy," in Murat Aydede, ed., *Pain: New Essays on Its Nature and the Methodology of Its Study* (Cambridge, MA: MIT Press, 2005), 373.

[117] Antonio R. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. B. Ponto, J. Parvizi, and R. D. Hichwa, "Subcortical and Cortical Brain Activity During the Feeling of Self-Generated Emotions," *Nature Neuroscience* 3 (2000): 1049-56; and M. Liotti and J. Panksepp, "Imaging Human Emotions and Affective Feelings: Implications for Biological Psychiatry," in Jaak Panksepp, ed., *Textbook of Biological Psychiatry* (New York: Wiley, 2004), 33-74.

[118] Jaak Panksepp, "Social Pain," 374.

[119] Jaak Panksepp, "Social Pain," 374. Panksepp cites P. D. MacLean, *The Triune Brain in Evolution* (New York: Plenum, 1990); Jaak Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions* (London: Oxford University Press, 1998); N. I. Eisenberger and M. D. Lieberman, "Why Rejection Hurts: A Common Neural Alarm System for Physical and Social Pain," *Trends in Cognitive Sciences* 8 (2004): 294-300; and K. D. Williams, *Ostracism: The Power of Silence* (New York: Guilford Press, 2001).

The associations between pain and emotional distress even show up chemically. The opiate analgesics which I already mentioned as having the effect of making people "not care" about their pain turn out also to be "quintessentially effective" in removing signs of separation distress in many different animals, including primates.[120] There are also chemicals naturally produced by the brain which reduce both pain and separation distress: namely, oxytocin and endorphins.[121] All of these chemical and brain-regional correlations between pain and feelings of sadness and emotional distress suggest the possibility of a phenomenal similarity among these experiences. I propose that their phenomenal similarity is their shared quality of badness.

There is also evidence for all our positive experiences' sharing a common phenomenal quality, one which is the exact opposite of that shared by negative normative experiences. Research conducted by Michel Cabanac and others shows not only that subjects are able to rate various pleasures and displeasures on a common scale but that their subsequent behavioral choices tend to produce the highest algebraic sum of pleasure as defined by these ratings.[122] The experiments of Cabanac and others asked subjects to rate the pleasure or displeasure produced by two different stimuli in

[120] Jaak Panksepp, "Social Pain," 376; J. Panksepp, B. H. Herman, T. Villberg, P. Bishop, and F. G. DeEskinazi, "Endogenous Opioids and Social Behavior," *Neuroscience and Biobehavioral Reviews* 4 (1980): 473-87; and J. Panksepp, S. M. Siviy, and L. A. Normansell, "Brain Opioids and Social Emotions," in M. Reite and T. Fields, eds., *The Psychobiology of Attachment and Separation* (New York: Academic Press, 1985), 3-49.
[121] Jaak Panksepp, "Social Pain," 376; Panksepp, Herman, et al.; and Panksepp, Siviy, and Normansell.
[122] For a summary of this research and a discussion of its theoretical implications, see Michel Cabanac, "Pleasure: the Common Currency," *Journal of Theoretical Biology* 155, no. 2 (1992): 173-200.

various combinations and then gave them opportunities to adjust the strength of one of the stimuli however they wished. One experiment paired sweetness and sourness.[123] Another, temperature and fatigue.[124] A third, chest fatigue and leg fatigue.[125] In these three situations, "the subjects' behavior were repeatedly coherent: in the bi-dimensional sensory situations imposed by the experimenters, the subjects described maps of bi-dimensional pleasure in sessions where their pleasure was explored, and tended to move to the areas of maximal pleasure in these maps, in sessions where their behavior was explored."[126] More simply put, "in a situation of conflict of motivations, one can predict the future choice of the subject from the algebraic sum of affective ratings of pleasure and displeasure, given by the subject, to the conflicting motivations."[127]

D. J. McFarland and R. M. Sibly argued in a 1975 paper that the brain needs some common scale on which to rate the importance of various reasons for action before it can send behavioral instructions to the body through the "behavioral final common path."[128] The research cited above shows "that sensory pleasure fulfilled the conditions required of a common motivational currency, at least in the case of

[123] Michel Cabanac and C. Ferber, "Pleasure and preference in a two-dimensional sensory space," *Appetite* 8 (1987): 15-28.

[124] Michel Cabanac and J. LeBlanc, "Physiological conflict in humans: fatigue vs. cold discomfort," *American Journal of Physiology* 244, no. 5 (May 1983): R621–R628.

[125] Michel Cabanac, "Optimisation du comportement par la minimisation du déplaisir dans un espace sensoriel à deux dimensions," *Comptes rendus des séances de l'Académie des sciences* 300, no. III (Paris, 1985): 607-10.

[126] Cabanac, "Pleasure: the Common Currency," 181-2.

[127] Cabanac, "Pleasure: the Common Currency," 182.

[128] D. J. McFarland and R. M. Sibly, "The behavioral final common path," *Philosophical Transactions of the Royal Society* (Series B), 270 (1975): 265-93.

behaviors selected which have clear physiological implications."[129] But in addition, further experiments confirmed the same general pattern when subjects were asked to compare the displeasure of being cold with the pleasure of playing a videogame,[130] or to compare physical pain with the pleasure of receiving money,[131] and then were allowed to make behavioral choices trading off these pleasures and displeasures. Their decision-making again reflected a use of pleasure as a "common currency." That is, despite the very different natures of these experiences, subjects were able to rate the pleasure or displeasure of all of them on a common scale, and their behavior reflected a tendency to maximize the algebraic sum of pleasure as previously rated.

The mere fact that we are able to make consistent comparisons between pleasures and pains produced by very different stimuli does not prove, of course, that the basis for these comparisons is a similar phenomenal quality. For that proof, we ultimately have to turn to introspection. And there, things can get quite difficult. We have to deal with claims like that of Alan Fuchs, who insists that

> there is obviously no felt quality or sensation common to all of the experiences we enjoy, which would have had to have been the case if the enjoyment consisted in having the experience along with a sensation of pleasure. … Consider, for example, a pleasing stimulus, the sound of a rock-and-roll band, or the taste of a great wine. Simple

[129] Cabanac, "Pleasure: the Common Currency," 182.

[130] Cabanac, "La maximisation du plaisir, réponse à un conflit de motivations," *Comptes rendus des séances de l'Académie des sciences* 309, no. III (Paris, 1989): 397-402.

[131] Cabanac, "Money versus pain: experimental study of a conflict in humans," *Journal of the Experimental Analysis of Behavior* 46 (1986): 37-44.

> introspection reveals, even in these cases, no element of sensory experience common to them all.[132]

And Fuchs is far from alone in concluding from introspection on the diversity of pleasures that there is no single phenomenal quality that they share.[133]

But it is difficult to see what justification Fuchs or others could have for being so sure that nowhere amid all of the many qualia that vary from one pleasure to another is there a single common phenomenal element. Fuchs says that, if there were such a common element, it would have to be obvious, but I don't see why this should be the case. As a general rule, we tend to pay more attention to contrasts in our phenomenal experience than to similarities. Things that don't change, we tend to ignore, until someone points them out to us. You might think that once someone asks us to attend to the positive phenomenal quality common to all pleasures, we ought to be able to do it, if indeed such a common phenomenal quality exists. The experiments of Cabanac seem to show that we *are* able to attend to this quality and report its intensity consistently. If we doubt whether this dimension along which we are easily

[132] Alan E. Fuchs, "The Production of Pleasure by Stimulation of the Brain: An Alleged Conflict Between Science and Philosophy," *Philosophy and Phenomenological Research* 36, no. 4 (June 1976): 494-505, p. 495.

[133] See J. C. B. Gosling, *Pleasure and Desire: The Case for Hedonism Reviewed* (Oxford: Oxford University Press, 1969), 28-53; Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), 493; James Griffin, *Well-being: Its Meaning, Measurement and Moral Importance* (Oxford: Clarendon Press, 1986), 8; T. L. S. Sprigge, *The Rational Foundations of Ethics* (London and New York: Routledge & Kegan Paul, 1988), 130; L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford: Clarendon Press, 1996), 92-3; Fred Feldman, *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy* (Cambridge: Cambridge University Press, 1997), 8, 132; Elijah Millgram, *Practical Induction* (Cambridge, Mass.: Harvard University Press, 1997), 123; M. Bernstein, *On Moral Considerability: An Essay on Who Really Matters* (Oxford: Oxford University Press, 1998), 25; T. Carson, *Value and the Good Life* (Notre Dame, Ind.: University of Notre Dame Press, 2000), 13-14; and D. Sobel, "Varieties of Hedonism," *Journal of Social Philosophy* 33 (2002): 240-56, p. 241.

able to compare our experiences is truly a *phenomenal* dimension, however, perhaps this is because we are comparing it with other kinds of phenomenal experience from which it greatly differs. Normative qualia are certainly very different from color qualia, from shape qualia (like the "smoothness" and "jaggedness" reported by the patient of Weiskrantz and Warrington), from touch qualia, sound qualia, taste qualia, smell qualia, and the qualia of proprioception. They're also very different from the qualia of nociception. But think about the extent to which all of these qualia are different *from one another*. This does not prevent them from all being phenomenal qualities. Normative qualia could be quite different from all of these other qualities and still share with them the one property that all of them have: that of being phenomenal.

Shelly Kagan has suggested that opposition to the view that pleasure is phenomenal may be reduced if this is taken to mean not that pleasure is "a single kind of mental state or experience, or a single shared component of all pleasant experiences, or even a kind of component" but rather that pleasure is a "*dimension* along which experiences can vary."[134] Citing the influence of Leonard Katz, he proposes that pleasure could be understood by an analogy with volume. Just as volume may be best classified not as a separate component of sounds but as a dimension along

[134] Shelly Kagan, "The Limits of Well-Being," in Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul, eds., *The Good Life and the Human Good* (Cambridge: Cambridge University Press, 1992), 169-89, p. 172.

which the components of sounds can vary, so too pleasure may be best classified not as a separate experience or component of experience but as a dimension of these.

If this analogy makes the thesis that pleasure is phenomenal more palatable, then so much the better. In saying that pleasure is a phenomenal quality, I mean to leave open whether it is best understood as a "component" or as a "dimension" of experience. The important thing for the purposes of defending a robust moral realism is that pleasure is *phenomenal*, that it is in some way immediately present in our phenomenal experience.

It may be, however, that understanding normativity as a dimension along which our phenomenal experience can vary—even a dimension along which *all* phenomenal experience varies[135]—will make it easier to identify, introspectively, the phenomenal presence of normativity. Instead of looking for some sensation that appears in just a few particular situations, one will examine those aspects of one's phenomenal experience that are so pervasive that one is tempted not to think of them as being phenomenal at all. Perhaps when we begin thinking of our phenomenal experience in this way, it will become progressively more clear that normativity colors large swathes of our phenomenal landscape.

While in the end there is no way to *demonstrate* that all positive and all negative experiences share common phenomenal qualities, because we cannot display

[135] This is the hypothesis of Cabanac ("On the Origin of Consciousness, a Postulate and its Corollary," *Neuroscience and Biobehavioral Reviews* 20 [1996]: 33-40; and *La quête du plaisir: Etude sur le conflit des motivations* [Montréal: Liber, 1995], Ch. 5) and of Hilla Jacobson.

phenomenal experience for collective examination, in this section I have nevertheless shown that the great diversity of positive and negative experiences does not rule out their having some common phenomenal quality. I have also shown that chemical and brain-regional correlations between pain and feelings of sadness and emotional distress suggest the possibility of a phenomenal similarity among these experiences. And I've pointed out that the fact that subjects can rate diverse pleasures and displeasures on a common scale, and that their behavior reliably reflects these ratings, makes it plausible that there is more similarity among these experiences than many philosophers suggest. Ultimately, however, we must turn to introspection for confirmation (or disconfirmation) of the existence of normative phenomenal qualities. And there we should not be too quick to come to conclusions, since it may take us time to identify accurately all of the components, dimensions, and relations of our phenomenal experiences.

V. Objective normativity and non-intentionality

Yet even if one comes to the conclusion that all our positive experiences share some one phenomenal quality of positivity and all our negative experiences share a phenomenal quality of negativity, one might still wonder about the further claim I have made: that felt goodness and badness can form the basis for judgment-independent normativity. One might doubt this further claim due to a confusion about just what sort of judgment-independent normativity is being proposed. For instance, someone might think that normativity grounded in positive and negative phenomenal experience cannot be judgment-independent because different people experience positive and negative qualia in the presence of different objects. The goal of this section, however, is to explain in detail a point I made in Section II: that the objective goodness and badness of instantiations of normative qualia is not some goodness or badness that they supposedly confer on an object but the goodness and badness of the instantiations of the normative qualia themselves. Although instantiations of normative qualia are very often associated with other phenomenal qualities or with physical objects—as in the cases of pain and of disliking certain foods—what is intrinsically good or bad is the goodness or badness itself, not an object which causes it or a sensation which happens to be felt simultaneously with it. This is what I call the fundamentally *non-intentional* nature of the normativity of normative qualia. At its most basic level, the normativity of normative qualia does not represent the goodness or badness of anything else but rather characterizes the instantiations of the qualia themselves.

At the same time, the normativity of normative qualia does form the basis for our perceiving things besides normative qualia as good or bad. The first time one perceives a particular object, one may also experience a negative normative quale, perhaps for some arbitrary reason. This conjunction, especially if it is repeated, may cause one's brain to strongly associate the object with negative normative phenomenology, in such a way that one's brain produces this negative phenomenology automatically upon future experience of the object, and even upon merely thinking of it or imagining it. When we "perceive" an object as bad, we are simply taking the intrinsic badness of our phenomenal reaction to perceiving it or thinking of it to be a property of the object itself.

Antirealists, of course, want to deny this account of the nature of normative phenomenology. They do not want to explain people's mistaken belief in the objective badness of objects by showing how they have conflated the object with something that *is* objectively bad, because they do not want to allow that there is *anything* that is objectively bad. Thus an antirealist who has agreed with all of the claims I have made thus far about normative phenomenology—that it is phenomenology and that it is found across many different types of experience—will resist the foundation of an objective normativity on this phenomenology, and they may do this by claiming that the phenomenology is purely a phenomenology of *taking* something to be good or bad, and that it need not be objectively good or bad in itself. That is, the antirealist may propose that normative phenomenology, if it exists, is purely intentional—purely object-directed—and thus claim that, if the normativity that it purports to locate in its objects is illusory, then there is no remaining goodness or badness of the phenomenology itself on which to found a judgment-independent normative system. What I need to show, then, is why such a purely intentional account of normative phenomenology ought to be rejected.

Perhaps the best way to argue for the essentially non-intentional normativity of normative qualia is to cite the fact that we sometimes experience a normative quale without its being associated with any other sensations which we feel it to be "about." Consider what we might call general "emotional" states. We sometimes experience a free-floating happiness or sadness, for example, which are not felt to be "about" anything in particular. We simply find ourselves having an overall positive or negative feeling, a feeling of well-being or of being ill at ease. Often we can't identify the cause of such feelings, even after long reflection; much less does it seem that there is any intentional object the goodness or badness of which is immediately consciously "represented" by our feeling. Consider what P. W. Nathan reports in *The Oxford Companion to the Mind*: "Excitation of certain parts of the temporal lobes produces in the patient an intense fear; in other parts it causes a strong feeling of isolation, of loneliness; in other parts a feeling of disgust; and in others sorrow or strong depression. Stimulation of some parts causes a feeling of dread rather than of fear, a dread without object, the patient being unable to explain what it is he dreads."[136]

Consider, too, this description of the effects of electrical stimulation in certain limbic areas of the brain given by Murat Aydede:

> Whatever kind of pleasure the subjects are experiencing, in most of the cases, the pleasure is clearly "objectless": it is not directed to (or, caused by) certain thoughts or sensations proper. This is why, I think, the reports are usually of a feeling of immense well-being, euphoria, or

[136] P. W. Nathan, "Nervous System," in R. L. Gregory, ed., *The Oxford Companion to the Mind* (Oxford: Oxford University Press, 1987), 527.

> elation. This feeling is aroused almost suddenly five to fifteen seconds after the electrical stimulation is applied, even in the case of serious pathological depressives….[137]

A major part of Murat Aydede's thesis in the paper from which this excerpt comes is that there is no sensory quality to pleasure. He means by this not that there is no phenomenology of pleasure at all but that there is only *affective* phenomenology when it comes to pleasure, and no unique *sensory* phenomenology, such as the nociceptive sensations had in experiencing physical pain. According to Aydede—whose goal is not ethical but is rather merely to analyze the experiences of pleasure and pain—one can have a positive feeling without any further sensation which it is "about." There seems to be such a thing as non-intentional, affective phenomenology.

But one might doubt whether this non-intentional phenomenology could explain intentional normative phenomenology as well. The sort of mental association I appealed to above in explaining our perception of the badness of an object likely sounds very Humean, and this might make some readers skeptical. Yet while I think that the various aspects of our phenomenology may be capable of being bound together in more ways than the mere temporal simultaneity that Hume cites, I think there is something quite right in Hume's attempt to atomize mentality and describe complex psychological phenomena in terms of some basic elements and their relations. And in fact, Humean mental associations play an important role in modern

[137] Murat Aydede, "An Analysis of Pleasure Vis-à-Vis Pain," *Philosophy and Phenomenological Research* 61, no. 3 (November 2000), 555-56. Aydede cites R. Buck, *Human Motivation and Emotion* (Chichester, UK: John Wiley & Sons, Inc., 1976).

neuropsychology. According to one influential theory of how neural networks learn, first proposed by psychologist Donald Hebb in 1949, the simultaneous or temporally proximal firing of two neurons, one of which synapses on the other, increases the strength of their synaptic connection and thus the likelihood that they will fire together in the future. When you have millions of such neurons each synapsing on large numbers of the others—each synapse with a strength relative to the neurons' past history of common firing—what you get is a system which, when part of it is put into a particular state by an external stimulus, automatically produces many other patterns, simultaneously and sequentially, that reflect the states the system has previously been in or gone into when similar input was received. What you get is a brain that naturally associates based on states that it's been in at temporally proximal points in the past.[138]

I'm not qualified to produce a neuroscientific model of normative phenomenology, but what I want to suggest, based on the models of current neuroscientific research, is that such associative mechanisms could explain how we have feelings of liking or disliking something without appealing to a mysterious, irreducible "intentionality."[139] When we consciously like something, or approve of

[138] Hebb's theory as well as another are described in Glynn, *An Anatomy of Thought*, 253-4. For a description of a network that works according to Hebb's theory, see J. L. McLelland, D. E. Rumelhart, and G. E. Hinton, "The Appeal of Parallel Distributed Processing," in D. E. Rumelhart, J. L. McLelland, and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1 (Cambridge, Mass.: MIT Press, 1986), 3-44, pp. 33-7.

[139] For an interesting reversal of this direction of explanation, see Hilla Jacobson, whose current project is to explain the objectionability of the phenomenal state of pain not as a brute phenomenal fact but in terms of something she says is "independently known to be objectionable": having a representation of a bodily state as simultaneously bad and obtaining. That is, she wants to take the representational nature

something, or take pleasure in something, I believe it is because our brains produce a feeling of goodness when the object is perceived or reflected on. Our experience of the constant conjunction of the positive normative quale and the object then causes us to remark that the object itself is good. (And the ease with which such associations lead to moral pronouncements about the intrinsic goodness or badness of things other than instantiations of normative qualia forms the basis of much of our moral error. I will discuss this topic at length in the next chapter.)

In sum, normative phenomenology often comes to be associated with other properties or objects. This can lead us to assume that a feeling of goodness or badness must always be *about* something further. But normative phenomenology can stand all on its own, and it does not lose its intrinsic normative character in doing so. When normative phenomenology is isolated, as seems to happen in cases of electrical stimulation of certain limbic areas, its positive or negative nature stands out clearly as a property *of the phenomenology itself* and not of any intentional object. This means that, even if it is clear that our normative phenomenology cannot be taken as evidence of the objective goodness or badness of other objects, our normative phenomenology itself may nevertheless be objectively good or bad.

of pain as brute and explain the badness of having the representation (that is, the badness of the *pain*, not just the badness of the bodily state it represents as bad) in terms of it. Her arguments leave unclear, however, why on her view a representation of badness's obtaining should itself be bad. This fact seems better explained if it's the case that we represent something as bad by associating phenomenal badness with the thing. (Hilla Jacobson, "The Evaluative Structure of Pain," work in progress, 2008.)

But perhaps it is still not clear that normative phenomenology can be good or bad in an entirely judgment-independent way. One might note, for instance, that the quality of one's normative phenomenology actually depends on the judgments one makes about whether a certain action or state of affairs is good. For example, if one thinks that it's good for people to marry and have lots of children, then when one sees someone married with a large family, one's perception will be accompanied by the experience of a positive normative quale. On the other hand, if one thinks it's bad for people to marry and have lots of children, then when one sees the very same family, one will experience a negative normative quale. Surely this is judgment-dependence!

Actually, this example includes *both* judgment-dependent *and* judgment-independent normativity. What is judgment-dependent is whether a particular person will experience a positive or negative normative quale in response to seeing a married couple with several children in tow. What is judgment-independent is that experiencing a positive normative quale is good and experiencing a negative normative quale is bad. While a particular person's judgments or attitudes will determine whether he or she feels a positive or a negative quale, the positive or negative nature of the phenomenology is intrinsic to the phenomenology itself. If a particular person feels a negative quale, no one can experience *that same quale* and have it be positive rather than negative.

What this means for ethics is that, though states of affairs or actions that we observe may evoke in us different phenomenal responses, it remains objectively true that certain phenomenal experiences instantiate the property of goodness and other phenomenal experiences instantiate the property of badness. The judgment-independence of normativity consists in the fact that an instantiation of a normative quale has an intrinsic, qualitative nature that is good or bad. This is why understanding the difference between intrinsic and intentional normativity is so important. If one's experiences of normative qualia were understood primarily as being *about* the perceptions that cause them—for instance, about the couple with lots of kids—then, because it seems quite clear that the goodness of the things perceived changes according to one's reaction, it would appear that normative qualia were no evidence of objective, judgment-independent value. However, because when we examine the instantiations of normative qualia on their own, we find that they have an intrinsic qualitative nature, it then becomes clear that normative qualia *do* offer us evidence of judgment-independent value: the value that is present in the instantiations of the qualia themselves. When we stop trying to project the qualities of normative phenomenology onto perceptions with which they are merely associated, we realize that, far from being an illusion, judgment-independent value exists in the realm most immediate to us. Judgment-independent value exists as part of the very fabric of our mental life.

VI. Conclusion

In this chapter, I have attempted to point out the existence of phenomenology with an intrinsically normative quality by giving examples of such phenomenology within our everyday experiences. I have emphasized the intrinsic rather than intentional nature of the normativity of this phenomenology and have explained how its intrinsic normativity makes it possible that it is objectively normative, while remaining purely phenomenal. Even if the reader is not yet fully persuaded of the existence of normative phenomenology or of its non-intentional nature, I hope that the possibility nevertheless seems live enough that the details of a robust realism founded on the existence of such phenomenology will be of interest.

If we do experience the intrinsically, objectively normative qualia I have described, and normative facts are at bottom facts about this phenomenology, then we can begin to see how the criteria for a robust realism could be met. First, we have the potential for explaining the nature of our concept of goodness: the phenomenal quality of goodness could provide the basic content of that concept. When we ask ourselves what it is we mean by "goodness," we can turn to this basic phenomenal experience for the answer. Second, we have an understanding of how things in the world objectively satisfy our concept of goodness. They satisfy it by instantiating the same phenomenal property that gave us our concept.

Third, we have an explanation of the way in which we can come to *know* which things objectively satisfy our concept of goodness. We have immediate awareness of the intrinsic goodness of certain of our own experiences because their goodness is itself a part of their phenomenology. (Nevertheless, we may need to become more practiced at second-order reflection on the instantiation of these phenomenal qualities. We may not presently be as good at second-order reflection on these qualities as we are at reflection on visual or auditory qualia.) Our method of acquiring knowledge about the goodness of *others'* experiences is more complicated, since we cannot directly perceive others' phenomenology, and perhaps we cannot truly be said even to indirectly perceive it. Yet in the same way that we normally infer from the situation and behavior of others that they are experiencing visual or auditory phenomenology similar to what we would experience in the same conditions, we can infer that, when others claim to feel pain or pleasure, they are experiencing normative qualia similar to our own. The justification for skepticism about the positive and negative qualities of others' phenomenology is no stronger than that for skepticism about any other aspect of others' phenomenology.

The fourth criterion requires explaining why we sometimes make erroneous moral judgments. This is a complicated issue which I will take up in the next chapter, in a discussion of Moore's Open Question Argument.

CHAPTER 4

ANALYTIC DESCRIPTIVISM

In the last chapter, I argued that all of our positive experiences share a phenomenal quality of goodness and that all of our negative experiences share a phenomenal quality of badness, and that instantiations of these phenomenal qualities have their normative character independently of anyone's judgments about them, thus making them capable of founding a robust moral realism. My assertion that instantiations of these phenomenal qualities are normative is based purely on their qualitative nature, on their "raw feel." My claim that this raw feel is normative does not derive from any further argument. I do not try, for instance, to explain how normativity might supervene on this raw feel. Nor do I appeal to empirical data about the causal regulation of our use of normative terms. I believe the relationship between these phenomenal qualities and normativity to be much simpler. I claim that the normativity of these qualia is a conceptual truth, and that in fact our experience of them is what *gives* us our concept of normativity.

Establishing a conceptual connection between normativity and some one or more descriptive properties is actually the key to meeting the criteria for a robust realism. In particular, it is the key to meeting the third criterion: explaining how we can come to know which things objectively satisfy our concept of goodness.[140] If it is only the descriptive properties of things that we observe, and if our concept of normativity has no descriptive content, then it will be impossible for us to discover evidence of anything's satisfying our normative concept. We will still be able to study to which things we do as a matter of fact apply our concept, but we will have to give up finding any judgment-independent justification for our applications of it. On the other hand, if our concept of normativity has descriptive content—if our idea of intrinsic goodness, for instance, includes the phenomenal quality of goodness—then we have an empirical criterion by which we can determine that certain things satisfy

[140] It may be suggested that a sensibility theory like that of John McDowell is able to meet this criterion, but I believe sensibility theories, at least as they have been formulated in the past, are actually quasi-realist rather than realist. Sensibility theories do not achieve the level of judgment-independence required for realism, since, though they appeal to properties of objects which *merit* certain evaluative responses, the determination of which properties merit these responses is ultimately still based on facts about our evaluative responses themselves: on our second-order evaluation of our more immediate evaluative responses as "intelligible" from the moral point of view. McDowell writes, "there need be no basis for critical scrutiny of one ethical concept except others, so the necessary scrutiny does not involve stepping outside the point of view constituted by an ethical sensibility" (McDowell, "Projection and Truth in Ethics," in Stephen Darwall, Allan Gibbard, and Peter Railton, eds., *Moral Discourse and Practice: Some Philosophical Approaches* [Oxford: Oxford University Press, 1997], 221). For this reason, I class sensibility theory as a quasi-realist—and thus ultimately antirealist—position. And indeed, David Wiggins, another sensibility theorist, plainly refers to his view as subjectivist (in "A Sensible Subjectivism," in Darwall, Gibbard, and Railton, 227-44).

our concept and others don't. The things that satisfy our concept of intrinsic goodness are those that have the phenomenal quality that is part of the concept.

One of the goals of this chapter is to explain in detail how positing a conceptual connection between our normative concepts and phenomenal qualities enables my view to meet the criteria for a robust realism. But before I get to explaining the advantages of this view, there is an important objection to it that needs to be dealt with. Positing a conceptual connection between normative concepts and phenomenal properties makes the view a target of G. E. Moore's Open Question Argument against all descriptive analyses of normative concepts.[141] Moore's argument is usually characterized as refuting versions of "analytic naturalism," views that posit a conceptual connection between normative concepts and some natural property or properties. In fact there's nothing in Moore's arguments that would limit their scope to naturalistic analyses of normative concepts rather than descriptive analyses in general. Moore's argument is thus more accurately characterized as against "analytic descriptivism,"[142] and whether phenomenal properties are natural or not—a question I will refrain from taking up here—the view I'm proposing does come within the purview of Moore's criticisms.

[141] Moore, *Principia Ethica*, 5-21.

[142] Frank Jackson suggests the use of a very similar term—"analytical descriptivism"—but his definition of it is too narrow. He defines it as the doctrine that ethical sentences "are a priori equivalent to and analyzable in terms of nonmoral ones" ("Cognitivism, a priori deduction, and Moore," *Ethics* 113, no. 3 [April 2003]: 557-75, p. 558). It seems that it ought rather to be defined as the view that ethical sentences are *a priori* equivalent to and analyzable in terms of *descriptive* ones, leaving it open whether these latter are moral or nonmoral.

The Open Question Argument seems to be the primary reason that descriptive analysis of normative concepts is presently unpopular—so unpopular, in fact, that it's rarely even discussed except as historical background to contemporary metaethics.[143] Some philosophers have argued that Moore's argument can only show much less than Moore thought it could—for instance, that an analysis of normative concepts is unobvious (though possibly correct)[144]—and Moore himself later conceded that his "supposed proofs [that good was indefinable] were certainly fallacious."[145] Yet, in spite of whatever fallacies Moore may have committed, many philosophers continue to feel that there is some force in his Open Question Argument—for instance, that it offers some strong intuitive evidence that normative and descriptive concepts are mutually exclusive categories—and this discourages these philosophers from taking descriptive analyses of normative concepts seriously. Making a descriptive analysis of normative concepts plausible is thus going to require explaining away the strong intuition provoked by Moore's Open Question Argument.

[143] Jackson is an exception. He defends analytical descriptivism in *From Metaphysics to Ethics* (Oxford: Clarendon Press, 1998) as well as in "Cognitivism, a priori deduction, and Moore."

[144] See David Lewis, "Dispositional Theories of Value," *Proceedings of the Aristotelian Society* 63, supplementary volume (1989): 113-37, p. 130. See also Shafer-Landau, who says that Moore's argument is not knock-down, but does "create a burden of proof on the ethical naturalist" "of showing how two apparently different things—rightness and maximizing happiness, for instance—are really one and the same thing" (*Moral Realism*, 56-58, 67). And see Michael Smith, "Should We Believe in Emotivism?" in Graham Macdonald and Crispin Wright, eds., *Fact, Science and Morality: Essays on A. J. Ayer's* Language, Truth and Logic (London: Basil Blackwell, 1986), 293-94; or Smith's "Moral Realism," in Hugh LaFollette, ed., *Blackwell Guide to Ethical Theory* (Oxford: Blackwell, 2000), 15-37.

[145] Moore, "Is Goodness a Quality?" in *Philosophical Papers* (London: Allen & Unwin, 1959), 89-101, p. 89.

Thus, I will spend the first half of this chapter addressing the Open Question Argument. In the process, more of the details of my view will become clear, and I hope they will not only serve to neutralize the Open Question Argument but also to make my view as a whole more compelling. In the second half of the chapter, I will return to my discussion of the virtues of analytic descriptivism: particularly, the way in which it meets the four criteria for a robust realism that I set out in Chapter 2.

I. The Open Question Argument

Moore's argument relies on the intelligibility of asking, for any proposed definition of 'goodness' as property *P*, whether things with *P* are really good. "[W]hatever definition be offered," Moore writes, "it may be always asked, with significance, of the complex so defined, whether it is itself good."[146] Moore argues that the intelligibility of this question is significant because, if a correct definition of 'good' had been given, the question would not make sense. For instance, if 'good' just meant pleasant, then asking whether pleasure was really good would be like asking, "Are pleasant things really pleasant?" And yet when we ask, "Is pleasure really good?", we seem to be able to take this question quite seriously and perhaps find ourselves reflecting on it at length. That is, we find that it is an "open question." According to Moore, this would not be the case if 'pleasant' and 'good' just meant the same thing.

[146] Moore, *Principia Ethica*, 15.

Allan Gibbard presents Moore's argument in a slightly different way in his book *Thinking How to Live*.[147] Gibbard asks us to imagine one philosopher, called Désiré, who believes that 'good' just means desired and another philosopher, called Hedda, who believes that pleasure and pleasure alone is good. Gibbard asks us to imagine, furthermore, that Désiré thinks that things besides pleasure can be good. What Gibbard points out is that it seems that this belief of Désiré's contradicts Hedda's belief that pleasure alone is good. Yet if Désiré expresses his belief by saying "Not only pleasure is desired"—a statement that, on his view, is synonymous with "Not only pleasure is good"—Hedda can simply agree that, yes, not only pleasure is desired and yet still retain her belief that only pleasure is good. Why? Because Hedda does not believe that all that is desired is good. And this belief of hers seems to us perfectly coherent. Thus Gibbard thinks we must conclude that the concepts *good* and *desired* are not equivalent. If they were, the statement "Not all that is desired is good" would be incoherent, but it is not.

Gibbard's argument rests on the same basic principle as Moore's, but Gibbard makes the openness of the truth of a proposition appear even more clearly by asking whether its negation is coherent. The idea is that we know that the truth of the proposition "All that is desired is good" is an open question because the statement

[147] Gibbard, *Thinking How to Live*, Ch. 2, especially pp. 23-29. The argument has been taken up by many other philosophers in the meantime. See, for example, A. J. Ayer, *Language, Truth and Logic*, 2nd ed. (London: Gollancz, 1946; 19th impression, 1962), 105; and Donald H. Regan, "How to be a Moorean," *Ethics* 113, no. 3 (April 2003): 651-77.

"Not all that is desired is good" is coherent. We know that the first statement is not a conceptual truth because the second is not a conceptual falsehood. The fact that the truth of these statements is left undecided by the mere concepts involved is, according to Moore and Gibbard, conclusive evidence that the concepts *good* and *pleasant* are not equivalent.

Given that I believe there is a conceptual relation between goodness and the phenomenal quality of pleasantness, what do I have to say about the coherence of the statement "Not all that is pleasant is good"? Or, more specifically, about the coherence of the statement "This positive normative quale is not good"?[148] Many people contemplating these statements are going to take them to be coherent expressions of a certain moral point of view, one in which things besides pleasant experience can be valuable and pleasure itself might even have negative value. Many actual moral traditions teach exactly these beliefs. Doesn't their coherence mean that our concept of goodness must not bear any conceptual relation to the phenomenal experience of pleasure?

My response is not to dispute the principle underlying the Open Question Argument. I agree that, to the extent that two concepts are equivalent, denying that all things which satisfy one concept satisfy the other will be incoherent. The coherence test must be very carefully applied, however, in order to achieve reliable results. The

[148] I think 'pleasantness', of terms already existing in the English language, is the one that comes closest to designating exactly the positive normative phenomenal quality I've been talking about, and I will use it as a synonym. 'Pleasure', of course, will refer to the instantiation of this quality.

term ‘good’ can be used in many different senses and, on a simple application of the coherence test, the diversity and complexity of these senses can mask the underlying basic conceptual relation between goodness and positive normative phenomenology.

I propose to examine some of the statements whose coherence might be thought to prove a lack of conceptual relation between goodness and the phenomenal quality of pleasantness, and to show that, in these cases, which I take to be representative, there is an explanation for the seeming coherence that still allows for a basic conceptual connection between goodness and pleasantness. I will identify three major sources of this seeming coherence, based in our complex usage of the term ‘good’ and in the various associations we make with terms such as ‘pleasure’. I will then proceed to give examples of some statements phrased in such a way that the complexities in our use of ‘good’ do not interfere with the operation of the coherence test, or at least do so to a much lesser extent. With this interference greatly reduced, and with proper concentration on isolating the phenomenal quality in question from any associations we might make with it, I believe the denial of the goodness of pleasure will appear plausibly incoherent, thus providing a counterexample to Moore and Gibbard’s coherence claims and rendering the Open Question Argument ineffective as an objection to the view that goodness bears a conceptual relation to pleasantness.

To begin, I want to consider statements having to do with the pleasure associated with taste. A pleasurable tasting experience includes many phenomenal qualities of sweetness, saltiness, texture, scent, etc., but what's essential for our discussion here is that it also includes a phenomenal quality—pleasantness—which I say is conceptually related to goodness. Consider whether the following statements are coherent.

(1) Pleasant taste is not good.

(2) Pleasant taste is not always good.

I think many of us will have mixed feelings about the coherence of such statements. The first formulation some people may find hard to think coherent. How, after all, could a pleasant taste be bad? Isn't pleasure exactly what makes a taste good? Our confusion in regard to someone who declares that pleasant taste is not good is, I think, a preliminary (but certainly defeasible) point in favor of pleasure's being conceptually related to goodness.[149]

[149] It might be thought that this evidence is defeated by the recognition that, while producing pleasure may be what makes a *taste* good, it is not necessarily good *simpliciter*. Moral philosophers often draw a distinction between attributive goodness—that which makes something a good member of a certain class of things—and goodness *simpliciter*. For instance, being a good chair, a good hairdo, or a good criminal requires having properties that are not necessarily good *simpliciter*, i.e., not good in the absolute sense we take to be the sort of goodness morality deals with. The goodness of a pleasurable taste might be just this attributive sort of goodness—it might be what makes a *taste* good—but this doesn't mean pleasure is good in an unqualified sense that would provide a basis for moral realism.

However, I believe the case of pleasure had in tasting something is relevantly different from cases of attributive goodness. While one can reasonably ask why one ought to want a good chair or a good criminal, and the answer will have to do with certain further ends that are served by a chair or a criminal, and better served by a *good* chair or a *good* criminal, it makes much less sense to ask this question about a pleasant taste. We don't need to have a reason to have a taste in the first place, which then provides a reason for having a *good* taste rather than a bad one. Rather, we want the taste because of the pleasure it produces. The case of pleasant taste is thus unlike cases of attributive goodness, because the pleasure is the ultimate justification for desiring the tasting experience, not the other way around.

On the other hand, the second formulation—“Pleasant taste is not always good.”—seems more immediately coherent. Perhaps this is because the wording of the phrase seems to acknowledge the immediate goodness of pleasant taste but then imply that some broader conception of goodness is not always served by it. In hearing this statement, perhaps we think of the fact that many things which taste good can harm our bodies. For example, cough medicine may have a pleasant taste, but if that pleasant taste leads a child to drink large quantities of it, the end result will be decisively bad.

The fact that immediate goodness can often lead to badness in the long term produces the first crucial complexity in our use of the term ‘good’. ‘Good’ can be used to refer both to the immediate phenomenal quality of pleasure and to its *instrumental* value (as well as to the combination of the two). This is one of the reasons why assertions of the goodness of things besides the experience of pleasure are not always false on my view. While nothing besides a phenomenal experience can be *intrinsically* good if intrinsic goodness is a phenomenal quality, other things can be *instrumentally* good if they produce more positive than negative phenomenal experience in the long run. That is, they have the quality of “to-be-promoted-ness” not intrinsically, but because of their causal properties, because of their ability to produce other things that do intrinsically have the quality of “to-be-promoted-ness.” This is also why we can truthfully say that the mere fact that something feels good doesn’t mean it is good. This does not mean that intrinsic goodness is not a felt quality. It just means that not all immediate pleasure leads to pleasure in the long run. The *all-things-considered* goodness of something depends not just on its intrinsic qualities but also on its consequences.

Consider how this fact could explain the coherence of certain moral philosophies such as asceticism. One might think that the fact that ascetics can coherently say that pleasure is always bad, or that certain Eastern philosophies can coherently teach that all sensation, including pleasure, is bad and to be avoided, proves that pleasure has no conceptual relation to goodness. However, the statement "Pleasure is always bad" is coherent on my view because it does not explicitly deny the *pro tanto* goodness of pleasure (that is, it does not deny that pleasure is good *to some extent*). It merely implies that, if pleasure is *pro tanto* good, its goodness is always outweighed by some further consideration. This further consideration might be that denying ourselves pleasure is the only way to avoid feeling pain, if avoiding pain is a more important goal than feeling pleasure. But even if no further goal which outweighs the *pro tanto* goodness of pleasure is specified, the statement "Pleasure is always bad" is coherent simply because it is coherent to posit that *some* such goal exists.

However, if an ascetic goes a step farther and either explicitly denies the *pro tanto* goodness of pleasure or specifies that the goal he believes to outweigh the *pro tanto* goodness of pleasure is something which has nothing to do with normative phenomenal qualities, things become more complicated. If he denies the *pro tanto* goodness of pleasure, I believe he *is* saying something incoherent, and the reason for his making such a denial—and its seeming coherent—has to do with our tendency to mistake feelings we have *about* pleasure for intrinsic normative properties of it. I will discuss this sort of confusion a bit later.

If, on the other hand, the ascetic claims that the *pro tanto* goodness of pleasure is merely outweighed by the intrinsic goodness of some goal besides the promotion of pleasure or the avoidance of pain, he is making a different sort of mistake: attributing intrinsic goodness to something besides an experience with normative phenomenal qualities. On my view, nothing other than an experience that instantiates normative phenomenal qualities can be intrinsically good, because to be intrinsically good is just to have these normative phenomenal qualities. However, whether it's *coherent* to attribute intrinsic goodness to things besides normative phenomenal experience is a different question. Fortunately, it's one we don't have to answer here, because, regardless of whether such an attribution is coherent, it seems that—given that there are such things as phenomenal qualities of experience—people do make the mistake of attributing them to non-phenomenal objects. For instance, if it's true that there are phenomenal qualities of color, it seems likely that many people think of the green phenomenal quality of their experience of grass as a property of the grass itself, without realizing that the greenness they're attributing is a specifically phenomenal property.

Even if this is not what goes on in the case of color, however, I believe most people do make this sort of mistake about intrinsic goodness. Before they realize that intrinsic goodness is a phenomenal quality, they attribute it to all sorts of things besides phenomenal experience. They make many of these attributions just because, whenever they think of these other things, the phenomenal quality of goodness is also evoked in their minds, and they attribute this goodness directly to the objects with which it is associated. But after one realizes that intrinsic goodness is a phenomenal quality, one sees that these attributions of intrinsic goodness are strictly false and that the only normative properties things besides phenomenal experience can have are dispositional: dispositions to produce phenomenal experience which *is* intrinsically normative. The truth of this realization is not undermined by the fact that so many people have not yet had it and continue to say that things besides phenomenal experience are intrinsically good. The prevalence of this sort of confusion, however, does make attributing intrinsic goodness to things besides phenomenal experience seem coherent, whether or not it actually is. And this means that the seeming coherence of philosophies which attribute intrinsic goodness to ends other than pleasure is no proof that intrinsic goodness is not conceptually related to pleasantness.

So far, I have given one major reason why the statement "Not all that is pleasant is good" seems coherent despite a conceptual connection between goodness and pleasantness: this statement can be used to express the truth that not all that is intrinsically good is instrumentally good (and what one thinks is instrumentally good may be based on confused attributions of intrinsic goodness to things other than phenomenal experience). The statement “Not all that is pleasant is good” can be used to express another truth as well: the fact that certain pleasures are *signs of instrumentally bad dispositions*. Consider that the statement “Not all that is pleasant is good” might be used in negatively evaluating the pleasure someone takes in torturing another. This pleasure is the product of the person’s disposition to take pleasure in the agony of another, and such a disposition is extremely bad because the desire for pleasure will lead the person to seek out situations in which he can induce agony in others. The pleasure of the torturer is thus “bad” in the sense that it is evidence of this very bad disposition.

No doubt some readers will object to this explanation of why the pleasure of a torturer is bad. Some will think that the pleasure taken in torture is not just a sign of an instrumentally bad disposition but is also bad in itself. They may even claim that a torturer’s pleasure has absolutely *no* intrinsic relation to goodness but is *purely* bad. Doesn’t the coherence of this claim tell against a conceptual relation between pleasure and intrinsic goodness?

I believe such a claim only seems coherent. It seems coherent because the badness of the disposition that produces the torturer’s pleasure is so great that it crowds out our awareness of the *pro tanto* goodness of the pleasure. That is, when we are considering whether the pleasure of a torturer is intrinsically good, we are not focusing on the felt quality of the torturer’s pleasure. Instead, we’re focusing on the fact that he’s a *torturer*, and thus on the fact that he is inflicting *extreme pain*. To determine whether his pleasure is intrinsically good, we ought not to consider any of this context, because context is irrelevant to the intrinsic properties of his pleasure. While having a disposition to take pleasure in others’ suffering is of tremendous instrumental disvalue—in fact, it is one of the most disvaluable character traits possible, because of the way it puts one’s own interests at complete odds with those of others—the tremendous instrumental disvalue of this disposition does not prevent the pleasure it produces from having the same intrinsic quality of goodness had by other experiences of pleasure. But if we don’t focus on this intrinsic, felt quality of the torturer’s pleasure when we consider statements about it, we are not going to feel any incoherence in saying that his pleasure is purely bad. (Shortly, I will discuss what happens when we *do* focus on the intrinsic, felt quality of pleasure.)

I have now presented two reasons that the statement “Not all that is pleasant is good” is coherent despite a basic conceptual connection between pleasantness and goodness: ‘good’ can refer not only to the intrinsic value of pleasure but also to its instrumental value or to the instrumental value of a disposition which produces it. The instrumental value of pleasure or of a disposition which produces it (and which may produce other effects as well) is an empirical question, not decidable by mere reflection on the concepts involved. This gives the question of pleasure’s goodness an “open” feel, despite a conceptual connection between pleasure and *intrinsic* goodness.

It might be suggested that we modify the Open Question Argument to eliminate ambiguities in our use of ‘good’, specifying that we are discussing only intrinsic goodness. This will not automatically clear up confusions between instrumental and intrinsic goodness, however, since the confusions that cloud our evaluation of the intrinsic goodness of pleasure in situations of torture are just one example of a wider problem: that we can’t always immediately tell whether the value of a state of affairs is intrinsic or instrumental. There are several reasons for this. First of all, our generic term ‘good’ normally lumps intrinsic and instrumental goods together, and so we aren’t required in the course of everyday conversation to practice distinguishing them. But furthermore, the better acquainted we are with the instrumental value of something, the more the possession of that instrumentally valuable thing comes to give us immediate pleasure: first, the pleasure of looking forward to the further pleasures it will secure for us; then, in time, a self-contained pleasure that doesn’t depend on considering future benefits.[150] That is, repeated experience of the instrumental good of something causes that instrumental good to become so closely associated with the thing itself that we no longer have to think explicitly about its instrumentality in order to feel goodness when we possess it or merely imagine possessing it. Having or contemplating the instrumental thing

[150] Mill makes this point in *Utilitarianism* (1863), reprinted in Albert William Levi, ed., *The Six Great Humanistic Essays of John Stuart Mill* (New York: Washington Square Press, 1963), 241-308, pp. 280-81.

becomes itself a direct source of pleasure, blurring the distinction between intrinsic and instrumental goods.

The same thing happens with badness. Having or contemplating something which produces bad phenomenal experience can itself become a source of bad phenomenal experience. For example, we may not be able to think of the pleasure of a torturer without also feeling some empathy for his victim, and this may lead us to think that the badness of the torturer's pleasure is intrinsic rather than merely instrumental.

In addition to mistaking the instrumental properties of pleasure for intrinsic ones, however, we may attribute to pleasure as an intrinsic property the goodness or badness that our minds merely *associate* with pleasure because of our past experience. This is the third source of the feeling of coherence we have about the statement "Not all that is pleasant is good," and the last one I'll discuss.

The term 'pleasure' is able to evoke a huge array of associations for each of us, including memories of the various pleasurable experiences we have had in the past and our imagination of the many experiences we have been told by others are pleasurable. For some, the term 'pleasure' may evoke primarily thoughts of sexual pleasure. For others, it may evoke primarily memories of afternoons spent reading and sipping coffee in a café. For still others, the primary association may be with a priest or pastor who emphasized the sinfulness of indulging in sensual delights and the eternal punishment one could expect to reap from such activity. For all of us, the associative landscape of ‘pleasure’ includes much more than a simple positive phenomenal quality.

This complexity of our associative landscape causes great problems when we attempt an intuitive, pre-reflective evaluation of pleasure. If we were brought up in an environment where the most frequent references to pleasure were in contexts where its sinfulness and baseness were emphasized, our feelings about pleasure are going to be largely negative. The sentence “Pleasure is bad” is going to make a great deal of sense to us. Even those of us who weren’t raised in an environment where bodily pleasure was explicitly condemned sometimes associate certain sorts of pleasure with negative feelings like shame or embarrassment. Given all of these associations, it is no wonder that the question “Is pleasure good?” has an “open feel.”

But the intrinsic value of pleasure is not determined by the goodness or badness of the myriad associations one makes with it. The intrinsic value of pleasure has to do only with the phenomenal quality of the experience itself, and this phenomenal quality, even in the case of the pleasure taken in torturing, is not itself bad. It has the very same phenomenal quality had by pleasures which are felt by others when they are gardening, taking care of the sick, or accepting a Nobel prize. What matters in determining the intrinsic value of pleasure is not what feelings we have when we are thinking *about* pleasure, but the feeling we have that *is* pleasure. We can have this latter feeling without contemplating it and without even consciously labeling it “pleasure.” And if we manage to direct our attention to this particular feeling—all by itself, allowing no other associations to interfere—I believe we will not be able to help but recognize that this feeling is intrinsically good.

So, if we clarify the meanings of our statements involving normative concepts in order to do our best to clear up ambiguities and prevent equivocation between various sorts of instrumental goods and the most basic concept of intrinsic good, and if we focus our attention on the qualitative nature of the *experience* of pleasure rather than on the many thoughts we may have *about* pleasure or about its context, I think we will begin to find some counterexamples to Moore and Gibbard's coherence claims. For instance, the following statements, by using the terms 'intrinsically' and 'unqualifiedly', help to narrow the implications of the coherence test to the connection of the fundamental concepts involved, and I believe most readers will find these statements more plausibly incoherent than those we've previously examined. Consider:

(1) All experiences with the phenomenal quality of pleasantness are intrinsically, unqualifiedly bad.

(2) All experiences with the phenomenal quality of unpleasantness are intrinsically, unqualifiedly good.

Someone may insist that all that is needed to show that the phenomenal quality of pleasantness is not conceptually related to intrinsic goodness is to show that the statement "*Some* pleasant experiences are intrinsically bad" is coherent. What I have spent this section arguing, however, is that because of the complexities in the way we use the term 'good', and because of the way we have difficulty separating the associations we make with pleasure from the phenomenal quality of pleasantness itself, we cannot be sure that the *seeming* coherence of the statement “Some pleasant experiences are intrinsically bad” is a sign of its *actual* coherence. What I am doing in offering Statements 1 and 2 is providing two examples of statements in which these complications do not interfere, or do so to a much lesser extent. Even if there is only a suspicion of incoherence in these statements, that is more than there ought to be if Moore and Gibbard are right that goodness bears no conceptual relation to pleasure, and badness no conceptual relation to the unpleasantness of pain. These statements are thus plausible counterexamples to Moore and Gibbard’s claim that all statements predicating badness of pleasure or goodness of pain are coherent, and they cast doubt on the ability of the Open Question Argument to categorically dismiss claims of a conceptual connection between normative concepts and the phenomenal qualities had by the experiences of pleasure and pain.

Gibbard might say that what gives any intuitive plausibility to the incoherence of these statements is not their actual incoherence but rather the fact that norms of this kind seem absolutely crazy to us. For instance, he says about Philippa Foot’s example of “a man who insists that clasping one’s hands is good, and for no reason but that it’s the clasping of one’s hands” that the man is “not mixed up in his concepts [i.e., not saying anything *incoherent*]; he’s got crazy views on what to do and why.”[151] Gibbard

[151] Gibbard, 28. Foot’s example can be found in her paper “Moral Beliefs,” *Proceedings of the Aristotelian Society* 59 (1958-9): 83-104, p. 85.

might say that, if we as human beings agree on *anything* related to what to do, we agree that we should behave in such a way so as *not* always to seek pain and avoid pleasure. The goodness of pleasure and the badness of pain are very basic, deeply held values of ours. Gibbard might even say it is an unavoidable fact about ourselves as the kind of beings we are that we simply *cannot* endorse always seeking pain and avoiding pleasure. But he would add that this does not mean that it would be *incoherent* for someone to think pain is always to be sought and pleasure always avoided. It is simply impossible for us to think so, being the creatures we are.

But faced with someone who says that pain is intrinsically, unqualifiedly good, I believe our puzzlement amounts to more than just thinking the person is making bizarre choices about how to live his life. I believe we are also at a loss to figure out what he means by ‘good’. Of course, we do have at least one strategy for interpreting his use of the term: an expressivist one. We could understand a predication of ‘good’ as a mere expression of approval, for example. Perhaps someone could approve of promoting negative phenomenal experiences, and if that is all that is expressed by their saying, “Pain is good,” then the statement seems to be coherent. But I believe we have a concept of good more substantial than an expressivist one, and it is the application of *this* concept to negative phenomenal qualities that we cannot make sense of.

The term ‘good’ can be used not only to express approval but also to refer to a particular phenomenal quality which *justifies* approval. This more substantial concept of goodness derives its content from our experience of pleasure, and it is this second, qualitative meaning of goodness which is at the root of our confusion when someone says, “Pain is intrinsically, unqualifiedly good.” If one focuses on this qualitative concept of goodness when considering the statement “Pain and all other experiences with the phenomenal quality of unpleasantness are intrinsically, unqualifiedly good,” I believe it will be hard not to feel the incoherence.

In light of the explanation I’ve given for the ease with which we find the statement “Not all that is pleasant is good” coherent, and in light of the plausible incoherence of Statements 1 and 2 above, Moore’s Open Question Argument should no longer be thought to support a quick dismissal of a conceptual connection between normativity and certain phenomenal qualities. However, despite my disagreement with Moore’s denial of a conceptual connection between normativity and any descriptive properties, I want to conclude this section by emphasizing an important similarity between our views.

Moore also draws from the Open Question Argument the conclusion that normative concepts are not definable in terms of *non-normative* properties,[152] and with this claim I agree. In arguing that goodness and badness are qualities of phenomenal experience, I have not been attempting to establish a conceptual connection between goodness and badness and some non-normative properties. My thesis is rather that these qualities of phenomenal experience *are* normative properties. Their normativity

[152] Moore, *Principia Ethica*, 9-17.

is their felt, descriptive character, independent of any judgments we make about them, and this is why they provide the basis for moral realism. So while Moore is right that goodness has no conceptual connection to any non-normative property, the overly simple application of his open-question test has kept many philosophers from considering the possibility that goodness is nevertheless conceptually connected to a certain property of phenomenal experience that *is* normative.

II. The advantages of analytic descriptivism

With the Open Question Argument out of the way, I want to return to discussing the advantages of analytic descriptivism. My plan is to explain how giving a descriptive analysis of normative concepts—and in particular, an analysis in terms of phenomenal qualities—allows a view to fulfill each of the four criteria for a robust realism given in Chapter 2.

Criterion 1: An account of our concept of goodness.

A view which posits a conceptual connection between normative concepts and phenomenal qualities has a ready-made account of the content (and origin) of our normative concepts: our normative concepts receive their content from our experience of these phenomenal qualities. On my view, our experiences of the qualities of pleasantness and unpleasantness give us our basic concepts of things' being worth seeking or avoiding, of things' being such that they not only cause approval or disapproval, but justify these reactions. These basic normative concepts are *qualitative*, and we come to have them in much the same way that we come to have our other qualitative concepts, such as those of phenomenal redness or phenomenal saltiness.

Against the idea that there exists moral "experience" that is capable of giving rise to our moral concepts, Crispin Wright says,

> What I doubt is whether we can find anything of sufficient rawness in the phenomenology of moral judgement to give the notion of 'moral experience' any serious work to do. The question is whether there are modes of experience which should properly count as moral but which would be possible for a normal human subject who possessed as yet no moral concepts. It is hardly a completely perspicuous question, but it is also hard to see what motive there could be for returning a positive answer. Very small children, to whom we should hesitate to ascribe any concept of humour, will laugh at grimaces and other forms of clowning, and may harmlessly be described as finding them funny. What would be a comparable, pre-conceptual finding of moral value? Suppose such a child is distressed by the sight of a jockey whipping his horse. Should that count as a primitive sentiment of moral disapprobation? It should be obvious that the question is underdetermined. Perhaps the child is frightened by the thundering of the horse's hooves, or the jockey's mask, or feels himself threatened. What is necessary, if the sentiment is to count as moral, is that its cause be conceived *by the child* in a certain way, and that its causality be dependent on its being so conceived. It has to be the horse's presumed distress, conceived as such, and even perhaps some conception of the mercenary motives for its affliction, which causes the child's distress. So the suggestion is that there is no basis for describing an affective response as moral unless the subject gives evidence of the conceptual resources which would suffice to explain it as such.[153]

[153] Crispin Wright, "Moral Values, Projection and Secondary Qualities," *Proceedings of the Aristotelian Society* 62, suppl. vol. (1988): 1-26, pp. 12-13.

I believe that Wright makes the same mistake made by most philosophers who have discussed moral phenomenology and the possibility of moral "perception." He assumes that the most basic moral sentiment is going to be something along the lines of moral disapprobation.[154] He is right, I think, to say that moral disapprobation, in the way we normally think of it, would require a rather complex understanding of the relation between the horse's distress and the motives of the jockey whipping it, and it would be difficult to determine whether the child perceived this relation and felt disapproval toward it before already understanding the basic moral principles and relations involved. But where Wright goes wrong is in not noticing that experiences of normativity can be much more basic than this experience of moral disapprobation.

Wright says, "Perhaps the child is frightened by the thundering of the horse's hooves, or the jockey's mask, or feels himself threatened. What is necessary, if the sentiment is to count as moral, is that its cause be conceived *by the child* in a certain way…." But even if the child's feeling doesn't qualify as "moral disapprobation" unless its cause is conceived by the child in a certain way, this should not lead us to ignore the basic normativity that is already present in a child's fright at the thundering of a horse's hooves or in his feeling of being threatened. Normative phenomenology is a very general phenomenon. It is present not just when we disapprove of something in the "all-things-considered" way associated with specifically "moral" pronouncements,

[154] For other examples of moral phenomenologists who seem to make this mistake, see Maurice Mandelbaum, *The Phenomenology of Moral Experience*; and Terence Horgan and Mark Timmons, "Moral Phenomenology and Moral Theory," *Philosophical Issues* 15, Normativity (2005): 56-77.

but when we feel fright, or pain, or sadness: when we feel badness (or goodness) in a merely *pro tanto* sense. It is at this very basic level of our experience that the normative concepts begin their development, by receiving their core qualitative content. Experiencing fear, pain, sadness, pleasure, and happiness gives us an understanding of badness versus goodness, and, as our capacity for logic and abstract thought grows, we are able to develop from these very basic concepts the more complex concepts of instrumental goodness, of something's being good for one person but bad for another, and finally of all-things-considered goodness—Wright's "moral" goodness—which takes into account both something's intrinsic goodness and its instrumentality in producing intrinsic goodness for all other experiencing subjects.[155]

Criterion 2: An explanation of how things in the world objectively satisfy our concept of goodness.

This criterion requires an account of a judgment-independent connection between our concept of goodness and the things that are claimed to satisfy this concept. Versions of analytic descriptivism meet this requirement by positing that our concept of intrinsic goodness contains descriptive content, and that the things that satisfy this concept are just those things that fit this description. As I've just explained, the particular analytic descriptivist view I defend posits that the descriptive content of

[155] In the next chapter, I will explain and defend my account of how all-things-considered goodness is a function of the intrinsic goodness and badness of individual instantiations of normative qualia.

our concept of intrinsic goodness comes from experience of the phenomenal quality of pleasantness. This means that those things capable of objectively satisfying this concept are going to be other experiences of this same phenomenal quality.

My view also provides an account of how things that are not phenomenal—for instance, actions or physical states of affairs—could come to have a judgment-independent connection to this core descriptive content of our normative concepts. Though actions or physical states of affairs cannot themselves instantiate the phenomenal quality of intrinsic goodness, they may causally contribute to the instantiation of this quality. If they do, then they objectively satisfy the slightly more complex normative concept of instrumental goodness: conduciveness to the production of intrinsic goodness.

Of course, this way of meeting the second criterion depends on our normative concepts' actually having a core of descriptive content. Antirealists—and even some realists, such as Moore—have long insisted that this cannot be the case, that the normative and the descriptive form two mutually exclusive categories. They claim that no matter how thoroughly one describes the world, it is always a further question which of the properties described are good and which are bad. This insistence on the fact-value distinction goes back at least to Hume, who insisted that one cannot derive an "ought" from an "is."[156] Yet while I agree with Hume if he is merely saying that one cannot derive a normative statement from a non-normative one, I must disagree

[156] David Hume, *A Treatise of Human Nature* (1740), Book 3, Part 1, Section 1.

with anyone who says that a descriptive statement cannot *also* be a normative one, and this for the following reason: certain of our phenomenal experiences are such that an accurate description of them *must be normative*.

To see this, imagine that you are a scientist taking an inventory of all the phenomenology present in human experience. You've written down the qualities of experiencing various colors, sounds, and smells. But there are two distinct phenomenal qualities which you can't quite figure out how to describe. In the end, you realize that the only way to describe the one is to say that it is "good," and that you can only describe the other by saying it is "bad." You have to mention the normativity of the phenomenology simply in order to describe it accurately. Its normativity is part of what you have to describe. And so you realize that the qualities of these experiences are simultaneously normative *and* descriptive.

Against the idea that the normative and the descriptive could overlap, Blackburn writes,

> It is up to a subject whether he cares about any particular secondary property in any way. If morality consisted in the perception of qualities, there would be a theoretical space for a culture which perceived the properties perfectly, but paid no attention to them. But however it is precisely fixed, the practical nature of morality is clearly intrinsic to it, and there is not this theoretical space.[157]

Blackburn says that "the practical nature of morality is clearly intrinsic to it," that there is no room for someone to perceive a normative property and yet not care about

[157] Blackburn, "Errors and the Phenomenology of Value," in Ted Honderich, ed., *Morality and Objectivity: A Tribute to J. L. Mackie* (New York: Routledge, 1985), 15.

it. But he also says that, if one *could* perceive normative properties, one *would* have the option of paying them no attention; thus morality must not consist in the perception of normative properties. And yet what reason does Blackburn have for insisting that no qualities whatsoever could possibly be such that their perception demands one's attention and concern?

Surely if we *did* perceive normative qualities, such a demand of attention and concern is exactly what we would get, precisely because Blackburn is right that "the practical nature of morality is…intrinsic to it." The real difficulty seems to be a lack of imagination with regard to how perception and motivation might be linked. It seems to me that, if we look closely enough at our phenomenal experience, we find two very good candidates for properties the perception of which is intrinsically motivating. Experiences of pleasantness and unpleasantness do seem very consistently to motivate us to continue or discontinue them, respectively, even if this motivation is often overridden by some other. One preliminary objection to this view might be that we're often not motivated by our observations of the pleasantness or unpleasantness experienced by others, but these are cases in which we aren't directly perceiving these qualities—we're directly perceiving other things which lead us to *infer* their instantiation—and so these cases don't tell against the thesis that *direct* perception of these qualities is intrinsically motivating. If we restrict ourselves to considering cases in which we actually *feel* pleasantness or unpleasantness, the empirical thesis that, absent any other motivations, we will be motivated to continue or discontinue the experience seems quite plausible.

I’m not going to attempt to prove that the experiences of pleasantness and unpleasantness are always motivating. I simply want to point out that, when we focus on experience of these qualities in particular, the claim that there are no descriptive properties such that perception of them is intrinsically motivating loses a great deal of its intuitive support. These phenomenal qualities seem to provide a plausible counterexample to the mutual exclusivity of the descriptive and the intrinsically motivating, and they cannot be dismissed simply by reasserting the mutual exclusivity of these categories.

It seems to me that the empirical connection that exists between the experiences of pleasantness and unpleasantness and motivation, along with the normative “feel” of these phenomenal qualities, gives us reason to take seriously the hypothesis that the descriptive and the normative are not mutually exclusive and that phenomenology is the location of a descriptive-normative nexus. The advantages of discovering such a nexus are great: it allows us not only to explain where we get the fundamental content of our normative concepts, but also to assert that the concepts of intrinsic goodness and badness *necessarily, conceptually, analytically, by definition* apply to instantiations of certain descriptive properties.

Criterion 3: An explanation of the way in which we can come to know *which things objectively satisfy our concept of goodness.*

After the discussion of Criterion 2, there is not much that needs to be added to show that my version of analytic descriptivism meets this third criterion. Once we have established a *necessary, conceptual, analytic, definitional* link between our normative concepts and certain descriptive properties, we simply need to show that the descriptive properties in question are epistemically accessible to us. Because my view posits that the descriptive properties conceptually linked to our normative concepts are *phenomenal* properties, it is clear that these properties are epistemically accessible to us. They are the most directly epistemically accessible properties possible (for the individual experiencing them, at least). Our knowledge about which things objectively satisfy our concept of goodness thus has a firm basis in our immediate experience of the normativity of certain phenomenal states, combined with inferences we make about the phenomenal states of others (to be discussed in Chapter 5) and our knowledge of which things are causally conducive to producing positive phenomenal experience.

Criterion 4: An explanation for why we are often mistaken about what is good.

This last criterion may seem like the most difficult for analytic descriptivism to fulfill. After all, if there’s a conceptual connection between normative concepts and certain descriptive properties, how could we possibly get things wrong? However, we’ve already begun to answer this question, in discussing the Open Question Argument. The reasons we often get the answers to moral questions wrong are basically the same reasons many people have been misled into believing that the Open Question Argument disproves the conceptual connection between pleasantness and goodness: we have to contend with the complexity of our normative concepts (which include the concepts of intrinsic, instrumental, and all-things-considered goodness) and with the existence of a large number of strong but contingent associations we make with pleasure, associations that obscure pleasure’s core, *pro tanto* goodness. (In trying to answer moral questions correctly, we also have to contend, of course, with our self-interest: our desire to believe that what is good for us is also what is good all things considered.)

All people, I believe, despite the fact that they do not all presently conceive the situation in these terms, experience *pro tanto* normativity as part of their everyday phenomenology. And, whether they are aware of it or not, it is the way in which objects, actions, and ideas come to be mentally associated with their normative phenomenology that causes them to declare these other things good and bad and motivates them to promote or avoid them. That is, we apply our normative concepts primarily by way of association.

Now, to some degree, this largely subconscious system of association is going to succeed in giving us accurate normative beliefs. As I explained in Chapter 3, the associational system works by forging mental connections between experiences which occur together or which are linked by some process of thought, and each time the same sorts of experiences or thoughts occur together again, these connections are reinforced. Thus when certain actions consistently result in negative phenomenology, we very quickly come to call those actions “bad” and to avoid them, even if what is intrinsically bad is only the negative phenomenology associated with them. In fact, we are sometimes so quick to form associations that, if our first experience with something was bad enough, we won’t give it a second chance, even when our bad experience was simply a matter of luck and not an accurate reflection of our future prospects.[158]

This sort of misassociation is just one example of the many ways that our associative system can lead us astray, despite its being in general a very useful tool. Consider, too, that while it’s helpful for us to be able to generalize from experiences we’ve had to experiences we haven’t (since none of us can ever experience everything in the world), the result of our generalizations is normative judgments that are inevitably biased towards the particular environments in which we’ve lived and the particular experiences we’ve had in those environments. We are inevitably going to find ourselves in various sorts of moral disagreements with people who have had different experiences, and in fact the way that these disagreements are most easily

[158] For evidence of the way that illness experienced in association with a certain taste can produce an enduring negative association with the taste, see J. Garcia, P. S. Lasiter, F. Bermudez-Rattoni, and D. A. Deems, “A general theory of aversive learning,” *Annals of the New York Academy of Science* 443 (1985): 8-21.

resolved is not through argument—though argument can help—but through each party's experiencing what the other has experienced. Through new experience, we develop more refined and subtle patterns of association between the world and our normative phenomenology.

It also helps, I believe, to do some extended reflection on the difference between intrinsic and instrumental good. This includes metaethical and metaphysical reflection on just what sorts of things can actually have intrinsic value. Coming to realize that intrinsic value is found only in phenomenal experience opens the way to understanding how that value is objectively related to other things in the world and can lead us intentionally to seek out new information about how our minds and the rest of the world work, so as to make more accurate judgments about what is instrumentally valuable.

In the end, the mere fact that we derive our core concept of normativity from our experience of normative phenomenal qualities does not mean that we will automatically recognize this fact. The *pro tanto* normativity of normative qualia is so basic and also so obscured by our various mental associations that we cannot expect it to leap to our attention immediately. Many conceptual layers have to be carefully peeled away to reveal the crucial nugget of phenomenal experience in which the normative and the descriptive overlap. This excavation of the moral mental landscape is a difficult task, and one that I can only make a start on in this dissertation, but I believe it is the central, crucial task of the analytic descriptivist, and the work that will finally produce a robust realism.

There remain some loose ends, however, even in this outline of my basic approach. Though I've described how positive phenomenal experience is by definition intrinsically good, and how an instrumental good is anything that is conducive to the production of intrinsic goodness (or to the reduction of intrinsic badness), I have not yet described how the scattered *pro tanto* goodness and badness of various experiences in various minds can make it the case that there is an objective truth about which states of affairs and which actions are good all things considered. It was crucial to establish first a fundamental conceptual relation between our normative concepts and epistemically accessible properties, but assuming that this has now been done, we still have to understand how the normativity of individual instantiations of goodness and badness can give us direction in our daily choices about what to *do*. I now turn to this question.

CHAPTER 5

FROM *PRO TANTO* GOODNESS TO GOODNESS ALL THINGS CONSIDERED

Thus far I have explained how it is that there are individual phenomenal experiences that are intrinsically good or bad independently of anyone's judgments about them. I have also explained that objects, actions, or states of affairs may be instrumentally good or bad to the extent that they causally contribute to the production or prevention of an intrinsically normative phenomenal experience. What I have not yet explained, however, is how the goodness and badness of individual phenomenal experiences can make it the case that a state of affairs that includes many different such experiences is good or bad *as a whole*, or, furthermore, how an action that produces many different experiences at many different times (and for many different subjects) can be one that we ought or ought not to perform, *all things considered*.

I believe that the correct method for deriving conclusions about what we ought to do, all things considered, is actually quite straightforward. As far as theory goes, it is simply a matter of adding up the positive value of all of the experiences produced by an action across all subjects and subtracting the negative value. Once we have determined in this way the all-things-considered value of each of our possible courses of action, we should choose the one that produces the highest balance of positive over negative value.

However, I have no illusions that this approach will immediately appeal to everyone. This is the method infamously known as the "utilitarian calculus," and its results often seem at odds with our pre-theoretical moral judgments. Those who disagree (or at least believe they disagree[159]) with the results of the utilitarian calculus

[159] I don't think the demands of utilitarian morality, properly understood in all of their complexity, are as divergent from common moral judgments as is usually supposed. There are differences, to be sure, but I don't think these are in the areas critics usually focus on. I will address this issue in Part III of the dissertation, especially in Chapter 6.

are going to wonder whether there is not some *other* way in which a notion of all-things-considered goodness could be derived from the intrinsic value and disvalue of individual experiences. Why could we not say, for example, that, while in general it is better to have more pleasant experiences and less unpleasant ones, in certain cases the potential for more pleasure simply isn't relevant? Or why not say that no amount of pleasure can outweigh suffering that has reached a certain intensity? Or that the relative distribution of pleasant and unpleasant experience among individuals is sometimes more important than their total quantities? Why does the method for reaching conclusions about all-things-considered goodness have to be straightforwardly additive?

I believe the utilitarian approach is the proper one because it is the only one that is metaethically defensible. We have not finished with metaethics once we have discovered how intrinsic goodness is objectively instantiated in the natural world. This discovery provides us with a realism that extends only as far as facts about intrinsic, *pro tanto* goodness. A complete realist theory must also include facts about all-things-considered goodness, and it must explain how these facts, too, are judgment-independent and epistemically accessible. The realism we're looking for is not a realism that affirms the existence of empirical, judgment-independent intrinsic value but then allows intuition or mere personal preference to take over the task of determining how this intrinsic value ought ultimately to influence one's decisions. If we are to have a thoroughly realist view, we must only appeal to ethical principles that we can defend by pointing to the existence of judgment-independent normative facts.

I believe that if we restrict ourselves to principles with judgment-independent justification, we will end up endorsing a straightforward additive approach to determining which states of affairs are all-things-considered good. My reasoning here depends on the fact that normativity on my view is not based on anything as abstract as reasons or maxims. It is based on concrete value, instantiated in our phenomenal experience. Since value is a concrete existent, more value just is, quite literally, more value. If we bring more positive experience into the world, we add more value, just as if we bring a baby into the world, we add another human being. Addition is the norm in the natural world, and if we are to be justified in thinking that all-things-considered goodness has a different basis, I believe we need to have epistemic access to some judgment-independent normative fact in addition to the intrinsic goodness and badness of phenomenology, a fact that supports a deviation from the additive norm. It doesn't seem to me that we have access to any fact of this sort, and thus I believe we should accept that intrinsic value is additive, just like numbers of people, sheep, and elementary particles. It is one of the goals of this chapter to argue for this conclusion.

Another goal of this chapter is to address an even more basic question about all-things-considered goodness. Before one even considers whether various instantiations of intrinsic goodness should be added or multiplied or put through some more elaborate function, one might wonder whether an individual agent has any reason to take into consideration the normative experience of anyone but himself. That is, one might wonder whether there is such a thing as all-things-considered goodness at all. Might it not be the case that each agent's experiences of pleasantness and unpleasantness are normative for him, but there is nothing that is normative for all agents?

Since the question of whether the normativity of various agents' experiences can be combined in any way at all is more fundamental than the question of just how it ought to be combined, I will begin with the former. And once I have argued that various agents' normative qualia are actually normative for *all* agents, not just for those agents who experience them, I will return to the question of how their normativity combines. In the final substantive section of the chapter, I will address some practical difficulties with determining what is all-things-considered best: difficulties in determining the relative intensities of different instantiations of normative qualia.

I. Are others' qualia normative for me?

It's understandable to wonder whether normative qualia are normative in an agent-neutral sense or only in an agent-relative sense. After all, any particular instantiation of a normative quale is only experienced by a single individual. If the experience of the quale is limited to an individual, it seems conceivable that its normativity could also be so limited. We might add to this the observation that it is generally our own normative experience that is most effective in motivating us to action. Perhaps the fact that others' normative qualia are not so effective at motivating us is one sign that others' normative qualia are not normative for us. Or perhaps there is a necessary connection between normativity and motivation, such that, if something doesn't motivate us (and wouldn't motivate us even if we were fully rational and fully informed), then it cannot be normative for us.

While I admit that there are good reasons to consider the possibility that the normativity of normative qualia is agent-relative, in the end I believe the fact of the matter is that normative qualia are normative *tout court*, without a special relation to any particular individual. To argue for this conclusion, I will first appeal to two features of the intrinsic nature of normative phenomenology itself: (1) the fact that it seems to present itself as having agent-neutral value and (2) the fact that it does not contain a reference to any particular action or agent. I will then explain how normative implications for action nevertheless arise from normative qualia and argue that, because of the way that implications for action are generated, being the person who will experience a normative quale makes no *a priori* difference to one's obligation to produce or prevent it. After presenting these basic arguments for the agent-neutral normativity of normative qualia, I will address a couple of objections. The first objection has to do with whether facts about people's actual or rational motivations circumscribe their normative obligations. The second has to do with the possibility of there being judgment-independent agent-relative reasons, perhaps ones which affect the degree to which we are obligated by agent-neutral reasons. My response to this second objection will lead into a final argument for agent-neutral normativity, based on reductionism about personal identity.

1. Positive arguments

In the last two chapters, I have heavily emphasized that normativity is intrinsic to a certain type of feeling. I have argued that when we have normative qualia, our experience is characterized by intrinsic goodness or badness; that is, it is characterized by the feeling that the very experience we are having is such that it either ought to be happening or ought not to be happening (though the feeling itself is just a *feeling*, not an awareness of or assent to any particular propositional characterization of it). We might wonder, though, whether the experience is really best described as being one of agent-neutral value, as being of a blanket "ought-to-be-ness" or "ought-not-to-be-ness." Might my experiences have a positive or negative quality that specifically calls for *me* to continue them or to stop them, without being normative in an agent-neutral way?

While the possibility of an agent-relative normative quality seems conceivable, I don't believe this is the sort of quality we actually experience. Consider the experience of pain. It seems to me that, when we feel pain, we don't feel that its negative character is only a reason for *us* to avoid it. We don't feel that our pain is merely something *we* have reason to get rid of but which is no reason for anyone else to help us. It seems rather that we feel that if anyone else can do anything to help us get rid of the pain, then they ought to, all else being equal. We don’t take the fact that a pain can only be *felt* by one individual to mean that it has a normative claim only on the actions of that individual. The pain is bad in a way that makes a *pro tanto* claim on anyone who is in a position to help. In feeling it, we feel something which either should or should not be going on, period.

Similarly, when we experience pleasure, we seem not just to experience a desire that the experience continue or an imperative which orders *us* to prolong it if we can. Rather, we feel a sort of value which gives anyone who could promote it a *pro tanto* reason to do so. Even if, as a matter of fact, we think there are other goals which are more important than pleasure, the feeling of pleasure seems to be such that, given the possibility of creating one of two worlds that are alike in every respect except that one contains pleasure and the other doesn’t, one ought to create the world containing pleasure—even if one won’t enjoy the pleasure oneself. And though there are various reasons for which we don’t actually expect other people to look after our pleasure on a regular basis, we do tend to think that, in cases where it won’t cost them anything, others do have a reason to promote our pleasure: the fact that pleasure is in itself good. What we feel in feeling pleasure is value itself, something that, just by being part of the world, makes it better. Granted, it makes the world as a whole better by making *our* life better, but, by being part of the world, it contributes to its overall value. My pleasure is good for me, and only enjoyable by me, but nonetheless it is good independently of anyone's judgments about it, and in such a way that anyone who can promote it has a *pro tanto* reason to do so.

This argument appeals, of course, to introspection on the nature of these phenomenal qualities, and different people's introspection may lead them to different conclusions—conclusions perhaps unavoidably influenced by the metaethical theories they already accept. I certainly don't expect an appeal to introspection to settle the question of the agent-neutrality of normative qualia. But I do think that the fact that the intrinsic qualitative natures of pain and pleasure seem (at least to many of us) to be objectively bad and good in such a way that they give all agents reasons to act makes it at least plausible to think that their agent-neutrality is intrinsic to their qualitative natures.

Another reason to think that the normativity of normative qualia is agent-neutral is that the experience of normative qualia doesn't contain any reference to any particular actions that it requires or to any particular agents who are required to take them. A normative quale is perfectly simple. It has only one component, with one dimension: a degree of "ought-to-be-ness," which can range from extreme "ought-not-to-be-ness" to extreme "ought-to-be-ness," that is, from extreme badness to extreme goodness. A normative quale does not include, however, any particular instructions as to what ought to be *done* to produce or prevent such experiences in the future, and it is not characterized by any reference as to *who* ought to be acting so as to produce or prevent such experiences. While certain sorts of actions may normally follow the experience of normative qualia—because of the particular way our brains are wired—the concepts of these resulting actions are not present in the simple, intrinsic quality of normativity. Even the quite general idea of an action that would produce or prevent an instance of normative phenomenology is not present in the normative phenomenology itself. The concept of action—especially of goal-directed action—arrives much later in our mental development than our first experience of pain or pleasure; the phenomenal quality of normativity is more basic than any of the notions we have about what to do in light of it.

Nevertheless, normative qualia do have implications for action. The connection to action appears when we add to our knowledge of the intrinsic goodness and badness of normative qualia information about their causal relations to the rest of the world. It is when we start to realize that other objects, states, and events could be conducive to the production of normative qualia or could prevent it that we realize that the intrinsic qualitative nature of normative phenomenology has implications for action. Once we know what actions would produce and prevent normative qualia, we see that the intrinsic normative nature of those qualia extends to those actions, making them instrumentally good or bad because of their relation to intrinsic goodness and badness. But these implications are not present in the normative qualia themselves. Rather, it's simply an analytic truth that, to the extent that something promotes an intrinsic good, it is instrumentally good.

Given that the normativity of normative qualia extends to actions just because these actions are more or less capable of producing normative qualia, there is no reason to think that a particular normative quale would *a priori* have more implications for one agent than for another. The implications that normative qualia have for an agent depend on that agent's ability to affect those normative qualia, and we have no *a priori* reason to expect that one agent will have a greater potential to affect a particular normative quale than another agent, even if he is the one who experiences the normative quale. Very often the pain and pleasure we feel is highly dependent on the actions of others, and this means that the actions of others very often have instrumental goodness or badness that comes from the intrinsic goodness or badness of normative qualia they produce in us.

Thus, when we make the move from the intrinsic goodness or badness of normative qualia to considering what might be done to produce or prevent them, we have no reason to restrict the implications of their normativity to certain actions or certain agents. What is contained in the feel of the normative quale itself is the absolutely simple phenomenal quality of ought-to-be-ness or ought-not-to-be-ness, and this normativity extends to anything that is in a position to produce or prevent it. As a consequence, if two human beings are equally capable of causing a positive normative quale or preventing a negative one, then they have equal *pro tanto* reason to do so.

2. Objections

Those are the basic positive arguments supporting the agent-neutrality of the normativity of normative qualia. I now want to address some objections to this view.

First, I want to consider the objection that the normativity of normative qualia might be limited to those subjects who are potentially *motivated* by those normative qualia. Reasons internalists claim that something can be normative for an agent only if that agent would be motivated by it *given that he or she were fully rational and fully informed*. The attractiveness of this theory seems to stem in part from our desire to believe that, if we are not motivated to do what we have reason to do, this must either be because we aren't aware of all of the facts or because we've made a mistake in reasoning. It likely also stems from an inability to see what besides our actual motivations and their logical implications *could* provide us with reasons to act. My description of normative qualia in the last two chapters is intended to make it plausible that there could be another source of reasons: the intrinsic, felt goodness or badness of other people's normative qualia. But what about the first worry, that failure to be motivated to do what we have reason to do ought always to be attributable to a mistake in reasoning or a lack of knowledge? Does my view claim that, when I know what normative qualia my actions will produce in others (or in my future self), any failure to be motivated by this knowledge is a mistake in rationality?

To answer this question in a meaningful way, we need to know what constitutes a mistake in rationality. If being rational is simply understood as being sensitive to the reasons there are, then we can say without trouble that someone who isn't motivated by prospects for others' normative qualia is not rational, because he is not sensitive to the reasons that others' normative qualia actually give us. This very open definition of "rationality" is the sort Christine Korsgaard appeals to, and that is why she says that "[t]he internalism requirement is correct, but there is probably no moral theory that it excludes."[160] If what it is to be rational depends on what the moral facts are, then any moral theory can accept that all rational people are motivated by the moral facts. Such a claim is true by definition. All people who are motivated by the moral facts are motivated by the moral facts.

Other reasons internalists give more substantive definitions of rationality, however. On Bernard Williams' view, being rationally motivated is being motivated to act on those reasons reached by sound deliberation from one's "subjective motivational set." One's subjective motivational set includes one's desires but also such things as "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects…embodying [one's] commitments…."[161] Since there could very well be a person whose subjective motivational set does not include a

[160] Christine M. Korsgaard, "Skepticism about Practical Reason," *The Journal of Philosophy* 83, no. 1 (January 1986): 5-25, p. 23.

[161] Bernard Williams, "Internal and External Reasons," in Stephen Darwall, Alan Gibbard, and Peter Railton, eds., *Normative Discourse and Practice: Some Philosophical Approaches* (New York: Oxford University Press, 1997), 363-71, p. 366. See also his "Internal Reasons and the Obscurity of Blame," in *Making sense of humanity and other philosophical papers* (Cambridge: Cambridge University Press, 1995), 35-45.

concern for the suffering of others, nor any other motivation which would require such a concern, my view is not reasons internalist on Williams' definition of rationality.

This shouldn't be a surprise, however, since realism itself entails the denial of this sort of reasons internalism. By asserting that normative facts are judgment-independent, realism excludes a necessary connection between normativity and anyone's subjective motivational set. It posits instead that normative force originates from something besides motivation: in the case of my view, from the intrinsic felt nature of certain phenomenal qualities. To object to this view by insisting that it's impossible to have a source of normativity independent of motivation is just to beg the question of realism.

Williams himself does not simply beg the question. Rather, he argues that "there is great unclarity about what is meant" by those who affirm the existence of reasons that are unconnected to an agent's subjective motivational set.[162] I can certainly grant Williams that many theories of reasons externalism have been unclear, but I hope to have offered a more lucid account of just what external reasons might consist in and why they might have normative force that is not always motivating, even to an entirely rational agent. Interestingly, Williams himself suggests what it might mean to say that an agent has a reason to act in a way that is nonetheless not connected to his subjective motivational set by a sound deliberative route. Williams

[162] Williams, "Internal and External Reasons," 370.

says such a statement could mean “that things would be better if the agent so acted.”[163] This is in fact exactly what I propose it means, and I am not sure why Williams doesn’t take the fact that things would be better if an agent acted in a particular way to be any reason for that agent to do so, unless it is simply because he’s stipulated an internalist meaning for the term ‘reason’.

Rather than continuing to discuss skepticism about the possibility of external reasons, however, I want to turn to concerns about agent-neutrality that might be had even by those who accept the existence of judgment-independent normativity. I might accept that a person’s positive and negative qualia provide judgment-independent reasons for action and yet wonder whether these judgment-independent reasons might still only be reasons *for that person*. I might think that they are reasons for her to act *regardless of the contents of her subjective motivational set*, but that they are nonetheless not reasons for anyone else. It would seem that this could be true in two different ways. Either a person’s qualia could just inherently only be reasons for that person (contrary to what I argued in I.1), or they could be reasons for everyone but be blocked by other reasons that people have. Let’s consider this second possibility for a moment.

The fact that the world would be a better place if there were more positive qualia and less negative qualia all around might not make it the case that any one person has a responsibility to work towards this goal, *if* there are additional reasons

[163] Ibid.

that do not stem from the objective value or disvalue of certain states of affairs but instead speak directly to what agents ought or ought not to *do*. It might, for instance, be the case that each of us has a reason to pursue particular projects or goals which interest us, regardless of whether they have any objective value that would give *everyone* a reason to pursue them. Also, it might be that each of us has a reason *not* to do certain things—such as steal, lie, or kill—even if doing these things would bring about an objectively better state of affairs.[164] But the crucial question is whether we're justified in thinking that any additional, agent-relative reasons of this kind actually exist.

The simple answer is no. (The more complicated answer I'll get to in I.4.) The intrinsic goodness and badness of normative qualia is the only sort of normativity for which we have so far been able to give a metaphysical and epistemological story which justifies us in believing that it exists in a judgment-independent way. No one has yet come up with a plausible story about what sorts of judgment-independent facts facts that are directly about what agents ought to *do* could be, nor how we could come to know about them.

[164] I take my description of the two sorts of agent-relative reasons from Nagel, "The Limits of Objectivity," Tanner Lecture on Human Values (Brasenose College, Oxford University, May 4, 11, and 18, 1979), available from http://www.tannerlectures.utah.edu/lectures/documents/nagel80.pdf, 119-20. He labels them "autonomous" and "deontological" reasons, respectively. He adds, "I am not sure whether all these agent-relative reasons actually exist. The autonomous ones are fairly intelligible; but while the idea behind the deontological ones can, I think, be explained, it is an explanation which throws some doubt on their validity" (120).

So if we are going to be justified in believing that normative qualia are only normative for those persons whose qualia they are, it seems like it's going to have to be not because further, agent-relative reasons cancel out the agent-neutral normativity of normative qualia, but because normative qualia are just inherently normative only for those persons whose qualia they are. And this view faces a huge metaphysical hurdle: giving a normatively significant account of agent identity through time.

3. A negative argument from reductionism about personal identity

The evidence we currently have about the nature of persons gives us little reason to be optimistic about discovering facts about personal identity significant in the way that would be necessary to support a view according to which normative qualia are only normative for the agent who is going to experience them. While in daily life many of us tend to think of ourselves as distinct, enduring subjects who take in all of our experiences and direct all of our actions, the reality seems to be that, while our experiences (and actions) are themselves connected in various ways, there is no experiencing subject over and above these interconnected experiences. Personal identity seems to be constituted by various relations among experiences, rather than by the presence of some additional, enduring entity. Furthermore, the relevant relations among experiences come in degrees, so that in certain situations there is no objective basis for drawing a boundary line between cases of identity and non-identity.

To give an idea of the metaphysical difficulties here, let me summarize some arguments given by Derek Parfit in his book *Reasons and Persons*. (Those who are familiar with these arguments may wish to skip the next three pages.) Parfit argues at length that we ought to be "Reductionists" about personal identity, that we ought to believe that "the fact of a person's identity over time just consists in the holding of certain more particular facts."[165] Particularly, he says we ought to believe that

> [w]e are not separately existing entities, apart from our brains and bodies, and various interrelated physical and mental events. Our existence just involves the existence of our brains and bodies, and the doing of our deeds, and the thinking of our thoughts, and the occurrence of certain other physical and mental events. Our identity over time just involves (*a*) Relation R—psychological connectedness and/or psychological continuity—with the right kind of cause, provided (*b*) that this relation does not take a 'branching' form, holding between one person and two different people.[166]

Parfit contrasts this view with the idea that there is some "deep fact" about personal identity which could make it the case that, even in the absence of any differences in psychological connectedness or continuity between two cases, a relation of personal identity could hold in one case but not in the other.

[165] Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), 210.

[166] Ibid., 216. For Parfit, "psychological connectedness" is the holding of direct psychological connections. Direct psychological connections hold between two subjects if, for example, one of them has memories of the experiences of the other, or if they share beliefs or desires. "Psychological continuity" is the holding of overlapping chains of enough direct psychological connections to make it the case that there is personal identity through time. Parfit says we can't plausibly define precisely how many such connections are enough, but he says "we can claim that there is enough connectedness if the number of direct connections, over any day, is *at least half* the number that hold, over every day, in the lives of nearly every actual person" (206). For more detail, see pp. 205-6.

Parfit presents three main arguments for Reductionism about personal identity. In the first argument, Parfit asks us to imagine that a few of the cells of his brain and body are replaced by cells qualitatively identical to those which composed Greta Garbo at age 30.[167] Then he asks us to imagine a case which is the same except that a few more of his cells are replaced by Garbo-like ones. In fact, we are to imagine all of the various possible degrees of such replacement, arranged on a spectrum ranging from no replacement at all to complete replacement by Garbo-like cells. This gives us all of the possible degrees of physical and psychological continuity with Parfit as he actually is at present.

It seems clear that, in a case with no replacement, Parfit remains himself. And it also seems clear that, with complete replacement, he ceases to be Parfit. However, there seems to be no point along the spectrum at which a plausible line could be drawn between Parfit's remaining himself and his ceasing to exist. Though we are normally inclined to believe that there is always a deep difference between someone's being himself and his being someone else, there is no evidence for such a deep difference anywhere along this spectrum. Parfit says we should conclude from this that the Reductionist View is true and that, in the cases in the middle of the spectrum, it is simply an "empty question" whether the resulting person would be Parfit. We know all of the facts about what happens in these cases, but it just so happens that the facts

[167] Ibid., 236-43.

do not provide a determinate answer about personal identity, because personal identity is dependent on certain kinds of continuity and connectedness which admit of degree.

Parfit's second argument appeals to the existence of actual cases of human beings whose consciousness has been divided by severing the corpus callosum which normally connects the two hemispheres of the brain.[168] Where previously there was one stream of consciousness, after the operation, there are two, each of which is not aware of the sensations or thoughts of the other and each of which can communicate independently with the external world, by way of writing with the hand under the control of that hemisphere. Parfit suggests that we imagine being able to detach and reattach our hemispheres at will, so that when we desire to have the two sides of our brain work independently—on a physics problem, for example—we can divide them, and then, when we've finished, reunite them. Such a case seems theoretically possible. It does raise some difficult questions for the Non-Reductionist, however. When a person's mind is divided, do both streams of consciousness belong to him, or does only one of them, or do neither? If both belong to him, as seems the most natural answer, what explains the fact that the experiences of both streams are not united with each other? The existence of an enduring subject seems to derive a lot of its intuitive support from its ability to "explain" the unity of consciousness, but the separate unities of two streams of consciousness cannot be explained by relation to the one enduring subject which the Non-Reductionist assumes to exist. If facts about connections

[168] Ibid., 245-52.

between neurons in the brain are sufficient to explain why certain experiences are unified in one stream of consciousness and other experiences are unified in another stream, what reason is there to believe that there is also an entity which “owns” both of these distinct streams of consciousness? It also seems implausible to say that only one or the other stream of consciousness belongs to the original person or that neither of them belongs to him, since they seem equally related to his past self and, when his brain is reunited, he will remember the thoughts and perceptions of both streams. In addition, either of these last two hypotheses would have to suppose either that there can be unity of consciousness without its belonging to an enduring subject or that an additional subject or two comes into existence every time the person divides his brain. The Reductionist View, on which there is no subject metaphysically distinct from the relations between experiences, seems much more plausible.

Parfit’s third argument appeals to the existence of actual cases in which human beings are known to survive when half of their brains have been destroyed.[169] Building on the evidence provided by these actual cases, Parfit asks us to imagine that he is in an accident with his two triplet brothers. In the accident, each of the two brothers’ brains is irreparably injured, and Parfit’s body is irreparably injured, although his brain is left untouched. Since half of a person’s brain is viable, the doctors split Parfit’s healthy brain and transplant each half into one of the bodies of his brothers. The question is, which of the two resulting people is Parfit? Each is psychologically

[169] Ibid., 253-60.

continuous with Parfit, because each has half his brain. But is Parfit one of them, both, or neither? If only one of the halves of his brain had survived, it would seem clear that Parfit would have survived. Since *both* halves have survived, we hesitate to say either one is Parfit, since there seems no non-arbitrary way to choose between them. But it would seem equally odd to say that Parfit did not survive, given that what did survive—*both* halves of his brain—is even more than what would have survived in the case in which we would easily agree that Parfit had survived. Here again we seem to have a question of personal identity with no determinate answer. And rather than insisting that there *must* be a determinate answer, one determined by the presence of an unobservable enduring subject, it seems we ought to embrace the Reductionist View on which facts about psychological continuity and connectedness constitute all the facts about personal identity there are.

I believe Parfit's arguments give us excellent reason to embrace the Reductionist View about personal identity: i.e., to accept that there are no enduring subjects over and above experiences related by psychological continuity and/or connectedness. But I believe that embracing this view of the experiencing subject makes it much less plausible that normative phenomenal experiences are only normative for the subjects who will experience them, or that there are agent-relative reasons that can override the normativity of phenomenal experiences of other subjects. On the Reductionist View, there is no independent subject that endures from the time an intentional action is performed to the time the normative experiences which result are experienced. There is only a web of experiences, beliefs, and desires which are sustained through time to varying degrees.

Consider what the consequences would be were we to hold onto some notion of agent-relative normativity along with Reductionism about personal identity. In the Parfit/Garbo case, we would have to admit either (1) that there is an exact number of cell replacements that would make future normative experiences of the resulting individual non-normative for the decisions Parfit makes before the surgery or (2) that normativity itself comes in degrees, according to the degree to which the experiences, desires, and beliefs of future subjects resemble our own or have evolved from them in a gradual way. The first option seems indefensible, since we have no evidence that such a sharp borderline exists. The second option also seems metaphysically arbitrary. Whether a present web of experiences, beliefs, and desires bears strong similarities to future such webs does not seem relevant to determining whether the normative experiences in those future webs is normative for the decisions taken by this one. It also doesn't seem relevant whether a very dissimilar future web will have evolved from the present web in many gradual steps rather than all of a sudden. What reason do we have to think that experiences, desires, and beliefs which are external to the intrinsic normative nature of a normative phenomenal quality have anything to do with the scope of its normativity? Our other experiences, desires, and beliefs may affect which normative qualia we experience, but given that a particular positive normative quale has been produced, its feel has an "ought-to-be-ness" that is independent of any experiences, desires, and beliefs which accompany it. These latter are no more intrinsically connected to its normativity than is the color shirt worn by the subject experiencing it. Should we say that the closer the color of the shirt I'm now wearing to the color of the shirt "I" will be wearing in the future, the more strongly normative the normative experiences of my "future self" are for my current actions? Surely not. But because the goodness or badness of a normative quale is internal to the quale itself, any attempt to make its normativity relative to some external relation is going to be just as arbitrary.

4. Conclusion

However, even if all normativity is at bottom agent-neutral, we should acknowledge that we may sometimes be justified in reasoning in an agent-relative way. It may sometimes be right for us not to worry about the goodness or badness of certain states of affairs, if our not worrying about them actually brings about a better state of affairs overall. For example, while the suffering caused by a civil war is in itself bad and does give everyone a *pro tanto* reason to end it, it might be the case that my own relation to the perpetrators and victims of the war is such that my worrying about it and doing my best to end it would actually bring about less good than would my attending to other problems. The fact that we are not physically or mentally capable of attending to all of the potential pleasure and pain in the world makes it *instrumentally* important to concentrate our efforts where they are likely to do the most good. Our recognition of the practical need to limit our concern is no doubt part of the reason we find it intuitive to think that the intrinsic goodness or badness of a situation does not automatically give *us personally* a reason to be concerned about it. What our inquiry into the metaphysics of normativity seems to show is that all intrinsically good or bad states of affairs are nevertheless *pro tanto* normative for us. Thus, in the absence of judgment-independent agent-relative reasons, all legitimate limitations on our obligations to promote the positive normative qualia of others must stem from the instrumental benefits of focusing our concern on those states of affairs where it will do the most good. And it's simply an empirical question what sorts of limits on our concern will do the most good.

In the end, others' qualia are *pro tanto* normative for us simply in virtue of their intrinsic goodness or badness—the ought-to-be-ness or ought-not-to-be-ness that is the defining quality of these experiences—and this intrinsic normativity necessarily gives instrumental value to any action that will promote or prevent it, regardless of whose action it is. Because there are no other judgment-independent sources of normativity, the only properties of an agent relevant to determining what he has reason to do are those properties which determine his effectiveness in producing positive qualia and preventing negative. If the fact that he is not the agent who will experience the qualia does not reduce his effectiveness in producing or preventing them, then it does not justify his ignoring them.

III. Are goodness and badness additive?

Assuming that every pleasant or unpleasant experience we could produce for ourselves or for others has a *pro tanto* normative claim on us, we have one more major question to answer in order to determine what we ought to do all things considered: how do we combine all of these individual normative claims?

I'm going to call "additivity" the view that we ought to sum up all of the separate claims and perform the action that will result in the greatest total quantity of positive normative experience minus negative. As I mentioned in the introduction to this chapter, this is not a popular view. Many people insist that it's important to aim for an equal distribution of good experience among different individuals, or that it's more important to have a small population of people each with lots of good experience than a much larger population where each person has less, even if the total quantity of good experience is higher in the larger population. It's not always clear, however, to what extent objections to additivity are based on doubts about its metaphysical and epistemic justification and to what extent they arise from a distaste for its normative results (i.e., the fact that it ends up requiring, permitting, or forbidding certain sorts of actions). In Chapters 6 and 7, I will address concerns about the view's normative implications. Here, I want to explain the metaphysical and epistemic reasons for holding it.

In this section, I will present an argument that, in the absence of any judgment-independent normative facts besides the intrinsic goodness and badness of certain phenomenal experiences, the *pro tanto* normative claims of these experiences demand to be added, and thus demand that we promote the state of affairs with the greatest total of good experience minus bad. I'm going to argue that to take these facts about intrinsic value and use them to construct some other view about what we ought to promote all things considered is to ignore the way in which the intrinsic value of these experiences objectively requires that it be taken into account in decision-making. To accept some other view isn't like choosing a position on a question that doesn't have an objective answer; it's to choose a position that's objectively wrong and that leads us to perform actions which make the world an objectively worse place by not maximizing the production of good experience over bad.

The argument for additivity is essentially this: the normativity of good and bad experiences is additive because these experiences are concrete manifestations of normativity. This means, among other things, that their normativity isn't derived from independent reasons or principles. *The normativity of an experience that has the phenomenal quality of goodness is based on the goodness instantiated in that experience itself, not on any principle that tells us that we have reason to produce this sort of experience.*

This is a way that my view importantly differs from the views of other utilitarians such as Henry Sidgwick and Hare. They defend their utilitarianism by arguing that the principle of utility—that one ought to do what will produce the greatest quantity of whatever is intrinsically valuable—is the principle which best systematizes our moral intuitions and renders them properly universal.[170] If the normativity of promoting good phenomenal experience and preventing bad were due to a principle defended like this—on the basis of our intuitions—then it would make sense to argue that the correct principle could require aiming at something other than merely the greatest total of good experience minus bad. It might also require taking into account our intuitions about the value of equality.

But the normativity of experiences with the phenomenal properties of goodness and badness doesn't depend on any independent principle, however it might be defended. Their normativity is constituted by the value that's manifested in the experiences themselves, in their phenomenal qualities. Their normativity is constituted by the fact that instantiations of these qualities, all by themselves, are reasons to act. That is, the phenomenal goodness that an action would cause *is itself* reason for that action to be performed. It doesn't just *give* us a reason, in virtue of some further moral fact, like the truth of the principle of utility; it *is* a reason. And because an instantiation of the phenomenal quality of goodness *is, all by itself,* reason to perform an action that would bring it about, if more such instantiations are promoted by an action, then there's more reason to perform that action. Similarly, if more instantiations of badness are *prevented* by an action, there's more reason to perform that action. Because the normativity of these experiences is concrete—because it's purely a matter of their

[170] See Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Chicago: University of Chicago Press, 1907, 1962), 382-88; and Hare, *Moral Thinking: Its Level, Method, and Point* (Oxford: Clarendon Press, 1981), 114.

individually instantiating certain intrinsic properties—it adds up just the way any other concrete parts of the world do. In the same way that, if there are more instantiations of pencilness, there are more pencils, if there are more instantiations of normativity, there are more reasons.

That's the basic argument. What I want to do now is explain the reasoning behind it in more detail, by looking at a theory proposed by someone who *doesn't* think value is straightforwardly additive and explaining why the theory doesn't work when it's applied to the intrinsic value of good and bad phenomenal experiences. The theory that I want to look at is Parfit's. He argues that more of "whatever makes life worth living" is not always better, and he uses happiness as his primary example of a good that "makes life worth living." Now happiness could be understood as including more than just positive phenomenal experience. On some moral theories, a happy life might have to include the presence of some non-experiential things, like certain relationships. But positive phenomenal experience is an important part even of broader conceptions of happiness, and it seems Parfit is intending for his non-additivity claim to apply to good phenomenal experience, among other things. So analyzing his view will be instructive for our purposes.

1. Parfit's alternative proposal

In his book *Reasons and Persons*, Parfit rejects what he calls the "Impersonal Total Principle." This principle reads, "If other things are equal, the best outcome is the one in which there would be the greatest quantity of whatever makes life worth living."[171] This is a general principle of additivity, and, if we believe that good phenomenal experience is what makes life worth living, then rejecting this Impersonal Total Principle will mean rejecting the more specific principle, "If other things are equal, the best outcome is the one in which there would be the greatest quantity of good phenomenal experience (minus bad)."

Parfit rejects the Impersonal Total Principle because it implies what he calls "the Repugnant Conclusion": "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living."[172] Put in terms of my view, the idea is that, for any world of people each having a lot of good phenomenal experience, there's another world in which, although each person's life has only barely more good experience than bad, the total amount of good experience minus bad goes up, just because there are so many more people in the world. Parfit believes that the conclusion that the world with many more people each living a barely good life is objectively better than the world with fewer people who each have a better life is repugnant to our moral intuitions. And in view of this, he concludes that we should reject additivity and instead endorse some other principle for determining the overall

[171] Parfit, 387.
[172] Ibid., 388.

value of a world. He considers replacing additivity with a principle that allows that, after a certain point, more happiness had by additional people doesn't increase the value of a world, or does so only to a diminishing degree.

Parfit ultimately rejects this particular proposal, for reasons other than those I'm going to give here, but what I want to show is that there's a very basic problem with it that Parfit doesn't recognize, which is that we can't deny that additional lives with more good than bad experience always make a world better *without denying that good experience has intrinsic value*. I'm going to argue that Parfit's proposal's being true of happiness or of any other good is simply incompatible with that good's having intrinsic value.

2. Against Parfit's proposal

Let me start a detailed defense of this claim by making clear what I mean by "intrinsic" value. This is simply going to be a more refined definition of the "concrete" normativity I mentioned earlier. Understanding this definition is crucial to understanding why the value of good phenomenal experience—and of anything else intrinsically good by this definition—has to be additive.

By "intrinsic good," I mean a good that has some value dependent on *nothing but its intrinsic properties*. Philosophers have contrasted intrinsic goods with various other sorts of goods, depending on their purposes. Probably most often, intrinsic goods are contrasted with instrumental goods, which I define as goods that have some value dependent on the fact that they contribute to another good, or would contribute to it under certain conditions. This contribution could be a causal one, or it could be due to the fact that the instrumental good is a necessary part of an intrinsically good whole, or that it's a relatum in a relation that's intrinsically good. What's important is that an instrumental good has some value that depends on its contributing to some other good, and that this value is thus dependent on something other than the intrinsic properties of the instrumental good itself.

Korsgaard, in her essay "Two distinctions in goodness," argues that intrinsic goodness should not be thought of as the correlative of instrumental goodness.[173] She believes intrinsic goods should be contrasted with extrinsic goods, and instrumental goods contrasted with final goods: things which are good as ends. She says that, "If intrinsic is taken to be the opposite of instrumental, then it is under the influence of a theory: a theory according to which the two distinctions in goodness are the same, or amount to the same thing."[174] She introduces the further category of goods that are valuable as ends in themselves, but only extrinsically so, in order to classify those things that are good because rational human beings choose to pursue them as ends in themselves. In these cases, she says, "Value…does not travel from an end to a means but from a fully rational choice to its object." [175]

[173] Christine M. Korsgaard, "Two Distinctions in Goodness," in her *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 249-74.
[174] Ibid., 250.
[175] Ibid., 261.

I intend to contrast intrinsic goods with *both* instrumental goods *and* extrinsic goods of the kind defined by Korsgaard. Both instrumental and extrinsic goods depend for their value on something besides the intrinsic properties of the things themselves. And there's even a further sort of "good" with which intrinsic goods can be contrasted—one I mentioned earlier: a good whose value depends on the existence of a separate moral fact, like the truth of the principle of utility.

In this dissertation, I tend to talk as if the only two sorts of value are intrinsic and instrumental, but this is because of the specific metaethical context I'm focusing on. Because I'm concerned with determining what's judgment-independently good, I'm setting aside extrinsic goods. And I'm also setting aside goods whose value is dependent on normative facts besides the intrinsic value of phenomenal experience. The goal is to see what normative facts follow strictly from the intrinsic value of phenomenal experiences combined with purely non-normative facts. In particular, I want to know what facts about the intrinsic value of worlds as wholes follow from these facts.

So an intrinsic good is one that has some value that depends on nothing but its intrinsic properties—not on anyone's judgments about it, not on any further moral facts, and not on its contribution to some other good. I'm now going to present an argument that Parfit's proposal can't be true of purely intrinsic goods. Parfit proposes that it might be the case for a certain good, *G*, that the world is made better by *x* value units each time a unit of *G* is added, up until a certain point, at which additional units of *G* no longer increase the value of the world, or do so only by smaller and smaller proportions of *x*. Economists call the value of a unit of *G* added at a particular point in the accumulation process its *marginal value* at that point. It's widely accepted that most goods decrease in marginal value as the total quantity of the good rises, just as Parfit is speculating happens with *G*. However, the model of changing marginal value does not work for one class of goods: *purely intrinsic goods*. This is because, in order for the marginal value of a good to fall (or rise), some of its value must be understood in terms of the value of some other good. For instance, a fifth slice of pizza even for a formerly very hungry individual doesn't have nearly as much marginal value as the first slice. This is because the value of pizza is measured in terms of the satisfaction of hunger and gustatory pleasure, both of which normally cease to be enhanced at or before the fifth slice. The marginal value of pizza can change because pizza is an *instrumental* good; it serves some further end which each additional slice can promote to a greater or lesser degree.

But purely intrinsic goods can't change in marginal value, because their value is based entirely on their intrinsic properties. The value of purely intrinsic goods is not affected by anything else in the world, and thus *not by the amount of the good that already exists*. What is good is the thing itself, not some relation that it bears to other things of the same type or to some further good that they collectively promote. Since the value of a purely intrinsic good can't be affected by the amount of the good that already exists, and *G*'s value *is* affected in this way, *G* can't be a purely intrinsic good.

Change in marginal value is only possible to the extent that something has value that is *not* purely intrinsic. Thus the intrinsic value of a good—and, in particular, the intrinsic value of good phenomenal experience—does not change depending on the quantity of it that already exists. Every additional unit of an intrinsically valuable good always adds the same amount of intrinsic value to a world.

3. Objections to additivity

Now there may be some hesitancy in accepting this, especially if one's thinking about some of the goods other than phenomenal experience that are often called "intrinsic." This brings us to the first of four objections that I'll consider to my view. It might be objected that great art, for example, is intrinsically valuable, while at the same time there seems to be a limit to how *much* great art we care about having in a world. Billions of great paintings are not obviously better than several million, for example. Perhaps this is a counterexample to the view I've been defending.

I think there's a better explanation of what's going on in the case of art, however, and that is that art is not of truly *intrinsic* value, but only of *immediate* value. While we value it for its immediately evident qualities and not qualities which merely produce some further object which we can *then* immediately value, there's a very important difference between an object of immediate value and something of truly intrinsic value. An object of immediate value is not valuable without some possible external relation to another thing, however immediate that relation may be. In the case of art, the necessary relation is to a viewer. The fact that we think more art becomes less valuable after a certain point probably has something to do with the fact that, after this point, the extra art will not increase the amount of art actually *enjoyed* in a world. There is a limit to the amount of art any individual or even any society can take in and benefit from, and thus a limit to the amount of art valuable for a world to have.

So, though it's sometimes referred to as "intrinsically" valuable, art is really only valuable in virtue of its possible relationship to someone who enjoys it, and the same goes for many other goods, such as an espresso, a sunset, or a symphony. They may be *immediately* valuable to those who enjoy them, but if, without the possibility of anyone at all enjoying them, they lack value, then they aren't truly intrinsically valuable. If the marginal value of a good *ever* decreases to zero, this is a sure indication that its value is not truly intrinsic.

Now Parfit does seem to consider happiness an intrinsic good, because he doesn't mention anything that makes it good besides the fact that it's *happiness*. Nevertheless, he proposes that, past a certain point, more happiness in the form of more happy humans might not add to the value of a world, or at least might only add to the world's value to a diminishing degree. How might he think this is possible, in spite of my argument that additional units of intrinsic goods always have the same value as every other unit? He's suggested that "we cannot…exclude the possibility that some things which we have assumed to have this kind of intrinsic value really have value of a kind which is in one way different. This value isn't instrumental, but it may partly depend on the number of other things of this kind that are in existence."[176]

The problem with this proposal is that happiness, or, more specifically, the component of happiness that's the good phenomenal experience, seems to have precisely the sort of intrinsic value that *doesn't* depend on other things in the world. What convinces us that the experience is good is its felt, phenomenal character, and this doesn't change depending on how much other happiness is in the world (apart from causal effects, of course, but these should have already been included in the report of people's total happiness). The number of other happy lives, where it doesn't causally affect me, is completely irrelevant to the happiness—and thus the value—of the life I'm living.

We should note, however, that Parfit does distinguish several times in *Reasons and Persons* between people's lives being valuable *to them* and their increasing the value of a world as a whole. I think this distinction is key to understanding why Parfit thinks it's coherent to claim that goods may have intrinsic value even if additional quantities aren't valuable, though I don't think the strategy ultimately works.

Parfit *has* to distinguish between value to an individual and value to a world because, while in order to avoid the Repugnant Conclusion, he wants lives after a certain quantity to stop adding value to a world, he obviously can't deny that every extra life worth living is at least valuable to the person living it. His mistake is to think

[176] In correspondence.

that happiness has *another* sort of intrinsic value besides the value that's enjoyed in the experience itself. Granted, the happiness of one person may be instrumentally valuable in making others happier or in producing some other good. But happiness isn't *intrinsically* valuable in a way that isn't just its felt, experiential value. Particularly, happiness isn't valuable *to a world* in a way that isn't just equivalent to its experiential value within that world. Worlds don't have distinct interests to which the happiness of the individuals in them could either contribute or not. It's obviously silly to talk of things being valuable "to" a world, as if the world as a whole could enjoy things. But even if we don't talk in terms of things being valuable "to" a world, and only in terms of things "making the world more valuable," it's puzzling what fact about a world could make it the case that a particular number of happy people is required to give it optimal value. There doesn't seem to be any feature of worlds as wholes which could explain how their intrinsic value comes to differ from the total intrinsic value of the things that make up those worlds.

But perhaps a further objection will be raised: that there's no reason for addition to be the default method for arriving at the intrinsic value of a world as a whole. Why should straight summing be the method that doesn't require any justification? The answer is, because addition *is* the default method for *concrete* things. So if the value we're talking about is actually *in the experiences*—if their value is just their individual instantiations of certain intrinsic properties—then more of the experiences means more of the value. It's just like when you have a table with four corners, and there's a penny in each corner: there have to be at least four pennies on the table. Addition is the only way to go in figuring out the total number of pennies. There is no question of counting pennies after the first one as only half-pennies. Each one counts for exactly what it is: a whole penny. And it's the same with the value of happy human lives. If we agree that we have four separate human beings who each currently enjoy an experience of value *x*, then the total value enjoyed in the world containing those four people must currently be at least 4*x*.

The only way that summing the value of individual lives would *not* be the appropriate method for finding their contribution to the value of the world as a whole is if individual lives were only valuable insofar as they contributed to some other good which could be promoted to a greater or lesser extent by additional lives. *Instrumental* value, for example, is *not* necessarily additive. The instrumental value of a compound thing like a world is not necessarily the sum of the instrumental value of its parts because the parts may interact with each other and enhance or detract from their ability to produce some further good. Similarly, the instrumental value of a statue is not the sum of the instrumental value of its parts, because what makes a statue as a whole valuable is the way the parts are related to one another and the impressions that together they make on a viewer. But the intrinsic value of something can't be enhanced or detracted from by its changing relations to other things or to viewers because, being intrinsic, this value doesn't depend on external relations or interactions, just on the intrinsic properties of the valuable thing itself. And so, as the things which have the intrinsic value pile up, the value piles up as well.

The only judgment-independent value we ought to recognize as attaching to a world as a whole that is not the simple sum of all the intrinsic value the world contains is the world's *instrumental* value. A world that contains a population of a certain size, for instance, can be of instrumental value because individuals often benefit from the fact that other people exist. They enjoy their company, as well as the benefits of the division of labor, the greater number of scientific discoveries, etc. But with all these benefits there might also be a size beyond which individuals derive no further benefit from the existence of additional persons. Our intuitions about the value of "worlds as wholes" are, I think, largely based on the instrumental value we imagine those worlds would have for possible members of them. Since after a certain point a higher population doesn't seem more instrumentally useful to us, we intuit that it isn't more valuable.

Instrumental value, however, is not the proper final measure for determining which of two worlds it would be better to actualize. Parfit's descriptions of worlds A and Z are stipulated to include *all* the happiness of *all* their inhabitants. Thus any instrumental benefits of a certain population size have already been accounted for and reflected in Parfit's descriptions. The descriptions of worlds A and Z already include the additional happiness (or misery) that has resulted due to population size in each case. The only question left is: which world has more *intrinsic* value? And because intrinsic value can't be enhanced or detracted from by interaction with the rest of the world—all causal interactions having already been accounted for in calculating the amounts of intrinsic value had by individual things or persons—the intrinsic value of a world just consists in the sum of the intrinsic value of the things in that world.

Parfit may have been led to reject this straightforward view by some of the arguments he makes in earlier sections of *Reasons and Persons*, where he shows that acting merely on agent-relative and/or person-affecting reasons is in some cases self-defeating and, in other cases, means preferring consequences that we all agree are worse. Parfit concludes from this that "our reasons for acting should become *more impersonal*."[177] He thinks, for instance, that unless we accept an impersonal sense of value, we can't say that an earthquake that kills more people is worse than an earthquake that kills fewer (aside from instrumental effects). While it's bad for each individual that *he or she dies*, the fact that *more people die* is not worse for any one individual—it's just bad for more. From the fact that *more people's dying* is not worse for any one person, Parfit concludes that the sense of "worse" which we rightly apply to the larger earthquake is "impersonal." And once he has accepted the necessity of an impersonal sense of value, it's probably easier for him to assert that, though persons' lives are valuable *for them*, they may not increase the impersonal value of a world.

This last move is a serious mistake, however. The intrinsic value lost in each earthquake is just the total intrinsic value that the lives of those individuals killed

[177] Parfit, *Reasons and Persons*, 443.

would have had (as well as the intrinsic value they would have produced in others' lives). The only sense in which this value is "impersonal" is that it encompasses the value of *more than one* person, viz., of *all* persons involved. It doesn't alter the value attributed to individual lives based on how many others are lived. It takes the value that would be enjoyed by each individual and counts every additional life with such value as an equal, additional reason for actualizing one world over another. Accounting for the fact that many people's dying is worse than a few people's dying thus does not require postulating that a world has a value that's not just the total value of its parts.

5. Concluding argument

In conclusion, let me give one final formulation of the core argument I've been making against Parfit's non-additive proposal. Given the restricted metaethical context we're discussing, if we were to say that the value of additional good experiences is irrelevant to ethical decisions after a certain threshold has been reached, this would be to ignore the intrinsic normativity of those additional experiences. Unless we acknowledge that each additional qualitatively identical experience of goodness strengthens our reason to bring about a certain state of affairs to the *same* extent as all preceding such experiences, then we're ignoring the intrinsic qualitative nature of the additional experiences. Since it's only the intrinsic, qualitative nature of these experiences that gives us a reason to act in the first place, the reason each additional one gives us has to be equal to the reason given by any other qualitatively identical experience. And not only must it be *equal* to those other reasons, it must be taken into account as an *additional* reason, because it's independent of any of the reasons we already had. It's an additional reason brought into existence by the additional instantiation of the property of goodness. Thus if we deny that an action's producing more good experience gives us more reason to perform the action, we're denying that the intrinsic quality of the additional good experience is *independently* sufficient to make it a reason to act—denying, in effect, that instantiation of that quality is intrinsically good.

It might be suggested that denying that we ought to produce the greatest quantity of intrinsic value does not mean we have to deny the reason-giving character of any particular intrinsically valuable thing. We do not have to say that a particular intrinsically valuable thing doesn't give us as much reason to act as other qualitatively identical things. After all, the idea that there are certain valuable things which are the "last to be added" to a state of affairs is just part of our conceptualization of the situation. We could just as well have started our counting with *those* valuable things and then come to the conclusion that it was some other things whose value was superfluous. So it's not as though the theory is denying the equal intrinsic value of all qualitatively identical things; it's just that all of the valuable things *combined* still only give us as much reason to act as some smaller number of things.

The problem with this proposal is that, while it's true that all of the qualitatively identical, intrinsically valuable things in a particular situation are taken into account equally in this method, each of them still counts for less in a situation of superabundance than in a situation of scarcity. The inequality is not between the reason-givingness accorded to two qualitatively identical things in the same situation, but between the reason-givingness accorded to qualitatively identical things in *different* situations. If I am choosing between an action which will produce Parfit's world A and an action which will produce his world Z, and I believe that the greater amount of happiness in Z doesn't make Z as a whole any better than world A, this means that I am granting each unit of happiness in Z less value than I am granting the units of happiness in A. *I am taking each instance of happiness in Z to be less reason-giving than the each instance in A, despite their being qualitatively identical.*

The same line of reasoning explains why additivity can't be abandoned in order to favor certain distributions of experience among individuals. Many people object that utilitarianism doesn't take into account the importance of equality: the fact that a world with a lower quantity of good experience minus bad could be better, if the experiences were much more equally distributed. Philosophers such as Rawls and Nagel charge utilitarianism with not "taking seriously" "the distinction between persons."[178] They accuse utilitarianism of taking a principle that individuals rightly

[178] John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), 27; Thomas Nagel, *The Possibility of Altruism* (Princeton: Princeton University Press, 1970), 134.

use for making decisions about their own interests and using this principle to make decisions at the interpersonal level, where they say it's not applicable. For example, while it's perfectly rational for an individual to choose to endure some pain now for the sake of a greater pleasure later, Nagel says this kind of reasoning is not applicable to interpersonal cases because "[t]o sacrifice one individual life for another, or one individual's happiness for another's is very different from sacrificing one gratification for another within a *single* life."[179,180]

Of course, a utilitarian can readily agree that trade-offs made between lives will have different consequences from trade-offs made within a single life. When an individual chooses to endure pain now for a greater pleasure later, the fact that she can look forward to the future pleasure lessens the impact of her present suffering in a way that it wouldn't be lessened if the future pleasure was to be enjoyed by someone else. Also, if one person has already been through a lot more suffering than another, it's possible that some present burden would be much harder for the first person to bear, because of physical and psychological effects of the previous suffering. A utilitarian can readily acknowledge these sorts of distinctions between persons—ones with

[179] Nagel, *The Possibility of Altruism*, 138.

[180] On this topic, Rawls writes, "whereas the utilitarian extends to society the principle of choice for one man, justice as fairness, being a contract view, assumes that the principles of social choice, and so the principles of justice, are themselves the object of an original agreement. There is no reason to suppose that the principles which should regulate an association of men is simply an extension of the principle of choice for one man. On the contrary: if we assume that the correct regulative principle for anything depends on the nature of that thing, and that the plurality of distinct persons with separate systems of ends is an essential feature of human societies, we should not expect the principles of social choice to be utilitarian" (*A Theory of Justice*, 25).

consequences for the resulting intensities of good and bad experiences—and can take them into account in decision-making.

What a utilitarian can't accept—or rather, what a *hedonistic* utilitarian can't accept—is that, if allocating a burden or a benefit to one person rather than another makes no difference to the duration or intensity of the good or bad experiences involved, there could still be a reason to allocate it one way rather than the other. But on my view, this is not an arbitrary rule, and it's not a result of mistakenly applying an intrapersonal decision procedure to an interpersonal decision. The necessity of attending only to differences which produce a difference in the duration or intensity of good or bad experiences comes from the fact that the reason-givingness of these experiences is entirely dependent on their intrinsic properties, and that there's no other judgment-independent normativity in the world. In this context, to say that a world in which a pleasure is had by person *P* is better than a world in which a pleasure of identical intensity and duration is had by person *Q*, where no other normative experiences are affected, is to ignore the equal value of both pleasures.

If one nevertheless wants to defend the importance of equality in the distribution of intrinsic goods, one could do so by defending the importance of a separate moral principle which sometimes *overrides* or *counteracts* the intrinsic reason-givingness of additional good phenomenal experience. But in a moral theory that's based entirely on the intrinsic value of phenomenal experience, there is no such further principle. And our goal here was to determine whether, in the absence of any further principle, the intrinsic value of phenomenal experience gives us an answer as to what we ought to do in cases where our options include producing multiple normative experiences. I've argued that it does, that the intrinsic reason-giving nature of each additional good or bad experience is an additional reason to perform an action that would promote or prevent it, respectively.

But does this mean that, if we embrace the metaethical view I've described, we have to accept Parfit's Repugnant Conclusion? Given that it's so repugnant, might it not be better to reject the metaethical view instead?

I don't think our feelings about the Repugnant Conclusion should lead us to reject this view, for two reasons. First, even on this view, it would never actually be the case that we ought to produce many billions of people living only barely happy lives, rather than a smaller number of people living happier lives. The resources necessary to produce the huge population would actually produce more total happiness if they were concentrated on fewer individuals, just because it takes so many resources to produce and maintain each additional life, before we can start increasing its happiness. So, in practice, my view will never make it the case that we ought to produce a world of barely happy people rather than a world with fewer, much happier inhabitants.

Second, even if the metaethical view I've proposed has some counterintuitive implications—and it's likely that it does have some—this doesn't automatically mean we ought to reject it. If we're aiming for an epistemologically respectable realism, then we have to consider what reason we have to believe that our moral intuitions accurately reflect judgment-independent truths. One reason our intuitive feelings may not be good indicators of the objective value of a world in which huge numbers of people live lives with barely more good experience than bad is that our feelings are likely strongly affected by sympathy with individuals in such a world. Since any individual in this world is worse off than any individual in a world where everyone has a higher balance of good experience over bad, we have more negative feelings about the former world. But the fact that our intuitions about the value of these worlds is particularly sensitive to the experiences of individuals—and likely not adequately sensitive to the value of large numbers—does not mean that the objective value of these worlds is not always increased by the existence of more barely happy people.

III. Practical questions

I now want to turn to some practical difficulties with trying to act in accordance with the theory of all-things-considered goodness I have defended in this chapter. There seem to be three distinct difficulties that are nevertheless all rooted in the difficulty of determining the relative intensities of different experiences of normative qualia. First, there is the fact that each person's qualia are only directly observable by him, at least as far as we know. Science may one day find direct neurophysiological correlates to our phenomenal experience, and perhaps we will then be able to know the exact nature of the qualia someone is experiencing by inference from his neurophysiology, but until that day, each of us is the observer of only one set of qualia—our own—and it is not clear how, without being able to observe the qualia of another person, we could know which are more intense. We will have to rely on other people's *reports* of their qualia, and it seems likely that different people could report their qualia differently, so that a similarity or difference in report doesn't necessarily indicate a corresponding similarity or difference in the qualia which give rise to it.

The second difficulty is in comparing the intensities of normative qualia that are due to very different activities. For instance, there seems to be some difficulty in comparing the pleasure taken in drinking a good cup of coffee with the pleasure taken in gazing at a fine piece of art, in having sex, in watching one's child take a first step, or in contributing money to disaster relief. Some philosophers have gone so far as to say that these pleasures are so dissimilar as to be entirely incommensurable. I have argued in Chapter 3 that, despite all of the differences among our various positive and negative experiences, they share at least one quality: positivity or negativity. It is this common dimension which makes them commensurable. And scientific experiments bear out people's ability to rate various experiences of pleasure and pain on a single scale. Experiments also show that their behavior reflects an attempt to maximize the algebraic sum of pleasure on this scale.[181]

[181] See the discussion of Michel Cabanac's research in Chapter 3, Section IV.

Yet even if we accept the commensurability of various normative qualia, it does seem that there is some difficulty in knowing exactly *how much* of the pleasure taken in seeing one’s child take a first step would be required to equal the pleasure of a certain sexual experience. These pleasures may be commensurable, but it is hard to see how we could determine just what their relative intensities are, and without that information, it is hard to see how their theoretical commensurability can be of any practical use to us.

The third difficulty is more basic than the other two. It is a difficulty about making *any* comparisons of just how much more intense one quale is than another, even when both of the qualia in question are had by the same person in the course of similar activities. It seems clear that we are often able to say, accurately, that one sexual experience was more intensely pleasurable than another, but is it plausible that we could make an accurate assessment of just how many more times pleasurable it was? Twice as pleasurable? Ten times as pleasurable? 6.894 times as pleasurable? Whether something is twice as pleasurable or ten times as pleasurable as something else makes a big difference as to what we should pursue. If a certain pleasure A is only twice as intense as a certain pleasure B, then if pleasure B lasts three times as long, we should still choose it over pleasure A. However, if pleasure A is ten times as intense as pleasure B, it should only be given up for pleasure B if pleasure B lasts over ten times as long. If it were given up for a pleasure B which last only three times as long as pleasure A, the loss in objective value would be significant. But if we can’t be sure whether one pleasure is closer to 2 or 10 times as intense as another, that's the sort of loss we're going to be regularly incurring. In addition to the fact that this is a big practical problem, we might wonder whether we even want to accept that there *is* a fact of the matter about just how much more intense one pleasure is than another, if we don't seem able to judge it reliably, even when the pleasures in question are both ours.

I want to address this last worry first and insist that none of these difficulties should lead us to believe that there is no fact of the matter about the relative intensities of normative qualia. The fact that this matter is not fully decidable by introspection or observation does not mean that there is no truth about it. One might think that, if qualia are just those things that we know directly, without intermediary, then there couldn't be any facts about them that are not obvious to us. It may seem that, if something's not obvious, it can't really be a characteristic of our experience.

This view of our knowledge of our own qualia is too simplistic, however. It is one thing to be able to experience a certain quale; it is another to be able to recognize similarities and differences between qualia. This second capacity requires specific and well-developed powers of introspection. It requires, first of all, that we be able to have in our mind's eye two distinct normative qualia at one time. If we want to compare a normative quale we are currently experiencing with one we've experienced in the past, we will have to be able to bring to mind a vivid memory of the past normative quale, and we will have to be able to do this without distorting the past normative quale, and without changing the experience of the other, present normative quale. Then, we will have to be capable of *comparing* the two. It seems that our brains do have a capacity to "inspect" our qualia and to come to believe propositions about their relations, but this involves more sophisticated neural operations than merely experiencing the qualia, and it is not something that we believe all creatures capable of experiencing qualia possess. Given that making comparisons among qualia requires a capacity over and above the capacity merely to experience qualia, it should not be surprising if it turns out that the comparisons this additional capacity enables us to make are not as fine-grained as the differences we are able to experience. The distinction between our experience of normative qualia and our introspection about them opens up the possibility of a gap between the qualities that our phenomenology actually has and those qualities that we are able to reflect on and develop propositional beliefs about. This means that, even if we are not able accurately to judge how much more intense one normative quale is than another, it may very well be the case that we did *experience* one normative quale with an intensity a precise number of times greater than another. The limits of our introspective capacities are not necessarily the limits of our experiential capacities. Given this, I don't see any reason to think there's no fact of the matter as to how much more intense some of our normative qualia are than others.

Still, even if there's a fact of the matter, we might be very worried about the prospect of not being able to discover it. What's the use of believing in judgment-independent truths about which things are more valuable than others if we can't know what these truths are and so can't use them to guide our actions? Can I offer any hope that we will be able to discover the relative intensities of pleasures, including across persons and across activities?

I believe that we do have the ability to make useful estimates of the relative intensities of pleasures. None of the three difficulties described above is fatal for this ability. Let's take first the difficulty of interpersonal comparison. It is true that we cannot directly observe others' qualia and that, to gain information about them, we have to rely on people's reports and behavior (judging, for instance, by the lengths they will go to in order to continue or to stop a certain experience). But such second-hand information is information nevertheless. It is not as precise as direct observation of others' qualia would be, because it is at some remove from the qualia themselves, subject to various intermediate influences like the subject's tendency to stoicism, and her abilities for introspection and for describing what she introspects. An individual's reports of the intensity of her normative qualia will also necessarily be affected by the range of her past experience of normative qualia.

But we should remember two things. First, while there certainly are significant differences among human beings that could cause their behavioral reactions and verbal reports of exactly similar qualia to differ, these differences should not be overestimated. Our bodies, including our brains, are much more alike than different. Given all of the similarities among us that we *can* directly observe, it makes sense to infer that there will be a lot of similarity in the phenomenology that we cannot observe, especially when two people share a culture, had similar upbringings, and make almost identical reports of their phenomenology. It seems to me we have good reason to assume that, if two people make very similar reports of their phenomenology, their phenomenology is similar enough to provide a rough guide for decision-making, and, if they make widely different reports, their phenomenology is different enough for this difference to be taken into account.

The accuracy of this assumption seems to be born out by various empirical data. For example, people's reports of their own happiness correlate well with the independent ratings of friends or acquaintances, as well as with those of interviewers meeting them for the first time.[182] Also, while people report different levels of pain when exposed to the same physical stimulus, their reports are nevertheless closely correlated with levels of brain activity in a particular area of the cortex.[183] This seems to indicate that there are at least rough correlations between differences in people's reports of their phenomenology and objective differences in their feelings.

The second thing to remember is that, by averaging information we collect from a large number of people about the intensities of the normative qualia they experience in various situations, we can arrive at a guide to the average relative intensities of normative qualia experienced in each of these situations. We can also compare an individual's reports to the whole range of his reports in other situations

[182] Ed Diener and Eunkook M. Suh, "National Differences in Subjective Well-being," in Daniel Kahneman, Ed Diener, and Norbert Schwarz, eds., *Well-Being: The Foundations of Hedonic Psychology* (New York: Russell Sage Foundation, 1999), 434-50.

[183] Robert C. Coghill, John G. McHaffie, and Ye-Fen Yen, "Neural Correlates of Interindividual Differences in the Subjective Experience of Pain," *Proceedings of the National Academy of Sciences of the United States of America* 100, no. 14 (2003): 8538-42.

and see whether across the board he gives stronger or weaker reports of the intensity of his qualia than others. In the end, there are many tools available to us for compensating for what differences exist in people's reports of and reactions to their normative qualia. We shouldn't despair of being able to make any significant interpersonal comparisons.

But what about our difficulties in determining the relative intensities of our own normative qualia? Can we tell just how many times better sex is than a cup of coffee or a look at a beautiful painting? I think this is actually going to be very difficult to accomplish by introspection. I already described some of the possible limits on our introspective capabilities: the difficulty in correctly remembering a past quale, and the difficulty in not altering the nature of a present quale by the very act of introspection. These difficulties lead me to believe that the most accurate way to make comparisons even among our own experiences is to adopt the interpersonal method: a reliance on reports, behavior, and brain activity. That is, the most accurate way to compare the intensities of two experiences is not to compare the memory of one with the present occurrence of another, nor to compare two memories, but to compare either one's brain activity during the experiences or, in the absence of this sort of data, to compare the immediate *reactions* one had to the experiences, reactions which can be objectively recorded and preserved without change, unlike memories. One can compare verbal reports of one's experiences, including such reports as ratings of intensity on a scale of 1 to 10, for instance, and one can also compare the extent to which one's non-verbal behavior was affected by the experience: the extent to which one was distracted from one's normal routine, the degree to which one dwelt on the experience afterward, the lengths one went to in order to prolong or terminate the experience. These last measures especially lend themselves to quantification, and, using a variety of such measures (perhaps also accompanied by the results of introspection about one's memories, taken with a grain of salt), it seems one could develop a rather good estimate of just how much more intense some pain or pleasure experiences were than others. In any case, there seems to be no reason to despair of being able to glean *any* useful information in this area.

Indeed, everyone should hope this is the case, not just those who support a utilitarian moral theory. All plausible moral theories give some importance to promoting pleasure and preventing pain, and thus they are all going to want to be able to tell which of two states of affairs is better from the perspective of promoting pleasure and preventing pain, even if they also consider other factors in determining what one should do, all things considered. If we have no way of determining the relative intensities of different pleasures and pains, utilitarianism isn't the only theory up a creek.

Fortunately, it does seem that the information we have about relative intensities of normative qualia is significant enough to make directing our actions by it worthwhile. We may never be able to obtain the optimal balance of pleasure over pain, but we will certainly be able to do significantly better than if we had no information about the relative values of people's normative qualia at all. The inevitable imprecision in our information should not prevent us from doing what we can, which in the end is a great deal.

IV. Conclusion

The central message of this entire chapter is that normative qualia are normative intrinsically—in the strongest, most literal sense of the word. Instantiations of pleasantness and unpleasantness are not normative because of their relation to anything outside of them but rather because of their own internal character, their raw feel. If their raw feel doesn't change, then the reason that they give us to act does not change, no matter who is experiencing them, how many other such instantiations there are, or how many of these other instantiations have been experienced by the same subject. Every instantiation of pleasantness or unpleasantness that feels just as good or bad as another must be counted as equally normative with it.

This intrinsic, qualitative normativity of normative qualia may well be the only judgment-independent sort of normativity there is. No one has yet made a convincing case for there being other observable facts which are objectively normative the way facts about these qualia are. Specifically, no one has made a plausible case for there being observable facts that tell us that certain instantiations of pleasantness or unpleasantness are only normative for certain agents, or that pleasant experience only increases the value of a world up to a certain point, or only when it is distributed in a certain way. In the absence of evidence for such facts, I conclude that there is no objective basis for any understanding of all-things-considered goodness except for the one which takes into account all potential instantiations of normative qualia, and takes them all into account equally.

I understand, however, that such a conclusion is going to strike many people as undesirable. For one thing, most of us have very strong feelings about the importance of equality in distribution. If the most plausible realist theory doesn't justify such feelings, might it not be preferable to turn to an antirealism which would not force us to give up these commitments? In the end, I believe the goodness of pleasure and the badness of pain are too obviously objective to permit disregarding them for the sake of principles that have no metaphysical or epistemological justification. On the other hand, I think that once all of the practical implications of a qualia-based utilitarianism have been worked out, we will see that it actually justifies most of our central moral commitments. It will not justify them as first principles, of course, but only as the best means to the end of promoting the greatest balance of pleasant over unpleasant experience. Nonetheless, it will give them an objective, judgment-independent grounding which antirealism cannot offer. Since the extent to which the realist view I have outlined turns out to justify our present moral commitments will likely strongly affect the degree to which it is taken seriously as an alternative to antirealism, I believe we must now turn to discussing these practical implications of the view.

PART III

THE NORMATIVE ETHICAL DEFENSE

INTRODUCTION TO PART III

In Chapters 3 and 4, I defended a metaethical view whose central claim is that intrinsic goodness and badness are phenomenal qualities of our experience. In Chapter 5, I argued that this metaethical view directly implies facts about what we ought to do, all things considered. I argued that, from the fact that certain phenomenal experiences have the properties of intrinsic goodness and badness independently of our judgments about them, combined with the assumption that there are no other judgment-independently normative properties in the world, it follows that we ought to do what will produce the greatest possible balance of positive over negative phenomenal experience. That is, it follows that something we might call "hedonistic utilitarianism" is true.

The fact that this metaethical view implies a hedonistic utilitarian normative ethical theory will likely be taken by many to be an important—perhaps even decisive—strike against it. Maybe it will even be considered by some to be a *reductio ad absurdum* of the metaethical view. Those who take the hedonistic utilitarian implications of the view to count against it will likely do so because they have strong moral intuitions which clash with what they take to be the demands of hedonistic utilitarianism. Faced with the alternatives of giving up their strong moral intuitions or giving up on finding a metaphysically and epistemologically robust realism, many people will choose to give up the latter.

I hope to show in the next two chapters, however, that we have reason to believe that the practice of hedonistic utilitarianism is not so different from the demands of our common moral intuitions as many people think, and thus that we don't in fact face this dilemma. I intend to show how the robust realism for which I've argued can actually *justify* a large portion of our moral intuitions, giving them a metaphysical and epistemological foundation.

Determining what the practice of hedonistic utilitarianism looks like is not particularly easy, however. We know, of course, that according to hedonistic utilitarianism one ought to do whatever will produce the greatest possible balance of positive over negative phenomenal experience. What is not so easy to determine is whether the actions that will produce the greatest possible balance of positive over negative phenomenal experience are generally the same as or different from the actions that are required by our common moral intuitions in the situations in which we find ourselves.

I add the qualification "in the situations in which we find ourselves" because it *is* clear that there are plenty of *hypothetical* situations in which what is utility-maximizing differs from what accords with our moral intuitions. (From now on, I will use the term "utility-maximizing" as shorthand for "productive of the greatest possible balance of positive over negative phenomenal experience.") We can simply stipulate a situation in which murdering someone is the utility-maximizing thing to do. Let's say that murdering in our hypothetical situation provides a utility gain of one person enjoying pleasure for one additional hour, versus the best non-murdering possibility for action. Most of us will have moral intuitions that tell us that one hour of pleasure is never enough reason to murder someone. Thus the demands of hedonistic utilitarianism are at odds with our moral intuitions in this case. And we can produce as many cases stipulated in this way as we want.

However, it's unclear that the divergence of our intuitions from the demands of hedonistic utilitarianism in cases where this divergence is stipulated is relevant to our decision whether to embrace hedonistic utilitarianism as the correct normative ethical theory. Whether we believe this divergence is relevant or not will depend on what sort of relation we believe our intuitions to have to theory. Do we expect our intuitions to guide us correctly in every hypothetical case? Or do we expect them to be somewhat tailored to situations we actually confront (or have confronted in the past)? It seems reasonable to expect our intuitions to be most accurate in situations similar to those that have influenced their development. If we can show that the demands of hedonistic utilitarianism and our moral intuitions generally coincide in these situations, we will have gone a long way towards making hedonistic utilitarianism look like an acceptable normative ethical theory.

It is the goal of these last two chapters to make plausible the claim that the demands of hedonistic utilitarianism and the demands of our moral intuitions generally coincide in the situations we actually confront. I will not argue that they *always* coincide, even in actual cases. I don't think our moral intuitions are perfect at picking out actions which are utility-maximizing. I believe our intuitions continue to evolve as our environment evolves, and as we subject them to more critical reflection. Certain of our intuitions, such as those that lead us to believe that the suffering of animals is less intrinsically important than that of humans, clearly need to be reformed, in light of the best theory we have about the judgment-independent normative facts. So I am not going to argue that our moral intuitions *always* coincide with the demands of hedonistic utilitarianism, just that it's plausible that many of them do, even ones that are often taken to be at odds with utilitarianism, such as intuitions about respect for rights.

I've also specified that I'm going to argue that it's *plausible* that these intuitions coincide with the demands of hedonistic utilitarianism. I can offer nothing like a conclusive proof that this is true. As I said above, determining exactly what the demands of hedonistic utilitarianism are is not easy. This is simply because determining the consequences of any particular action is not easy. The results of our actions depend on an extraordinary number of different factors, and which actions are best in which situations is a very complicated empirical question that I obviously can't answer in anything like a conclusive way in the space of these two chapters. What I can do, however, is to point out some general patterns in the consequences of our actions, patterns which suggest that following our intuitions—in things like refraining from murder, keeping promises, and telling the truth—tends to produce the best consequences in the situations in which we actually find ourselves.

I am going to divide my discussion of practical objections to hedonistic utilitarianism into two chapters. Chapter 6 will focus on objections to the consequentialist aspect of the view. The general idea of this sort of objection is that our moral intuitions require us to perform or refrain from certain *types* of actions, even if the consequences of not doing so might be better in particular cases. In this chapter, I will argue that certain features of the situations we actually confront make it the case that the decision procedure that is optimal from a utilitarian point of view has this same "deontological" form. Thus the deontological nature of our moral intuitions is not a reason to reject utilitarianism. In Chapter 7, I will go on to address objections to the particularly *hedonistic* consequentialism supported by the metaethic of the previous chapters. I will show why the fact that we value things besides experiential states is not a reason to reject hedonistic utilitarianism, since it can provide a metaphysical and epistemological foundation for these values.

CHAPTER 6

THE PRACTICE OF UTILITARIANISM

Attempts to show that the demands of utilitarianism actually coincide with many of our deontological moral intuitions have a long tradition, going back at least to David Hume.[184] The great nineteenth-century utilitarians—Jeremy Bentham, John Stuart Mill, and Henry Sidgwick—all did extensive work in this area.[185] And in the

[184] David Hume, *A Treatise of Human Nature* (1740), Book 3; and *An Enquiry concerning the Principles of Morals* (1751).

[185] Jeremy Bentham, *Introduction to Principles of Morals and Legislation* (1789); John Stuart Mill, *Utilitarianism* (1863); Mill, *On Liberty* (1859); Henry Sidgwick, *The Methods of Ethics*, 7th ed. (Chicago: University of Chicago Press, 1907), Book IV, Chapter III.

twentieth century, the job was taken up by J. J. C. Smart,[186] Richard Brandt,[187] and R. M. Hare,[188] among others.

What is puzzling is that this work seems to have done little to change general attitudes toward utilitarianism. Non-utilitarians still seem generally to operate on the assumption that it is obvious that the demands of utilitarianism are at odds with respect for rights and other strong intuitive moral concerns of ours. Perhaps this is partly because they haven't clearly distinguished between the demands of utilitarianism in hypothetical and actual cases, as I argued we ought to do in the Introduction to Part III. If they haven't made this distinction, I suppose it's because it hasn't been demonstrated clearly enough in what ways actual cases differ systematically from the hypothetical ones taken to be evidence against utilitarianism. Or because, even where these differences are acknowledged, it isn't clear that they are great enough to make it the case that utilitarianism generally coincides with our moral intuitions in the actual cases. I hope to make both of these points clearer in this chapter.

[186] J. J. C. Smart, "An Outline of a System of Utilitarian Ethics," in J. J. C. Smart and Bernard Williams, *Utilitarianism: for and against* (Cambridge: Cambridge University Press, 1973), 1-74.

[187] Richard B. Brandt, *Morality, Utilitarianism, and Rights* (New York: Cambridge University Press, 1992), Chs. 7-11.

[188] R. M. Hare, *Moral Thinking: Its Levels, Method, and Point* (Oxford: Clarendon Press, 1981); "Ethical theory and utilitarianism," in H. D. Lewis, ed., *Contemporary British Philosophy IV* (London: Allen and Unwin, 1976), 113-131, reprinted in Amartya Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982), 23-38; and "Utility and Rights: Comment on David Lyons's Essay," in J. Roland Pennock and John W. Chapman, eds., *Ethics, Economics and the Law*, *Nomos* 24 (New York: New York University Press, 1982): 148-57.

It may also be that some people find it difficult not to take seriously the persistently non-utilitarian intuitions they have about hypothetical cases. I think it's perfectly natural that the feelings our actual experiences have ingrained into us get extended to hypothetical cases, and that these feelings stay very strong even if we doubt whether they indicate the truth about what we ought to do. The fact that these feelings are strong and persistent, however, shouldn't lead us to take them as accurate guides to normative facts in situations very different from those that influenced their development.

I take as representative of contemporary dissatisfaction with utilitarian justifications for rule-following the complaints of David Lyons and Bernard Williams. Lyons argues in his papers "Utility as a Possible Ground of Rights" and "Utility and Rights" that utilitarianism is "hostile to the idea of moral rights."[189] He also argues that this hostility to the idea of moral rights undermines the normative force of legal and other institutional rights. Using the mundane example of parking in front of someone's driveway on a crowded city street, Lyons argues that there are many conceivable ways in which it could be utility-maximizing to violate someone's right to have their driveway clear, many of which nonetheless do not persuade us that we are permitted to violate this right. He argues that this is just one of many possible

[189] David Lyons, "Utility and Rights," in J. Roland Pennock and John W. Chapman, eds., *Ethics, Economics, and the Law*, *Nomos* 24 (New York: New York University Press, 1982): 107-38, p. 107. See also Lyons, "Utility as a Possible Ground of Rights," *Noûs* 14, no. 1, 1980 A.P.A. Western Division Meetings (March 1980): 17-28.

examples that show that we understand rights as imposing *constraints* on utility-maximization, an aspect of our moral intuitions that utilitarianism is simply unable to accommodate.

Utilitarian responses to such arguments generally point out forgotten utility considerations that are supposed to make it the case that respecting rights is utility-maximizing in more cases than originally thought, showing that the idea that we actually place constraints on utility-maximization is just an illusion. About these forgotten utility considerations, Bernard Williams writes,

> The certainty that attaches to these hypotheses about possible effects is usually pretty low; in some cases, indeed, the hypothesis invoked is so implausible that it would scarcely pass if it were not being used to deliver the respectable moral answer, as in the standard fantasy that one of the effects of one's telling a particular lie is to weaken the disposition of the world at large to tell the truth.[190]

While Lyons seems most concerned with the theoretical impossibility of utilitarianism's supporting constraints on utility-maximization, Williams is more concerned about the details of utilitarian arguments that such constraints are unnecessary.

It seems to me, however, that neither Williams nor Lyons has fully understood the way in which utilitarian arguments for respecting rights or for following other general rules are intended to work. For one thing, the likely weakening of certain generally beneficial social practices that occurs when one acts against them—in telling

[190] Bernard Williams, "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, *Utilitarianism: for and against* (Cambridge: Cambridge University Press, 1973), 75-150, p. 100.

a lie, for instance—is just one factor among many that generally make acting against them disutilitous. There are many other factors which, when taken together, provide strong utilitarian reason to make almost all of our decisions according to very general rules, like those of truth-telling, promise-keeping, and respecting rights to life, limb, and property. Furthermore, utilitarianism *can* justify constraints on *conscious* utility-maximization, since certain epistemic and motivational limitations of human beings make it the case that we often maximize utility by consciously employing a non-consequentialist decision procedure.

In this chapter, I examine in a systematic way the utilitarian justification for employing a decision procedure other than the conscious calculation of expected utilities. I present arguments that conforming to the Principle of Utility ("Do what maximizes utility.") actually requires us to make most of our decisions according to fairly general rules, such as keeping promises, telling the truth, and respecting rights such as those not to be killed or interfered with in certain ways.

In Section I, I explain why the metaethical theory I've defended does not allow us simply to declare ourselves rule utilitarians, defining what we ought to do as that which conforms to the rules that would have the best consequences if universally followed, or if followed by a certain subset of persons. I explain that, if following rules more general than the Principle of Utility is to be justifiable on the basis of the metaethical view I've presented, it's going to have to be because, *in the individual* *situation we're considering*, following these rules is the procedure most likely to lead us to conform to the Principle of Utility.

In Section II, I show, in general outline, why it might be more utilitous to make decisions according to a set of general rules than by employing what I call the "Straightforward Utilitarian Decision Procedure," that is, by actually considering the many consequences of one's actions and choosing the action with the highest expected utility. In Sections III through V, I discuss three features of actual situations and the particular reasons they give us to make decisions according to general rules. Section III addresses uncertainty, Section IV addresses the need we have to coordinate our actions with others, and Section V addresses limitations on human motivation that make it the case that we promote the best consequences if we refrain to some degree from interfering in others' lives. In Section V, I also discuss the way that our self-interested bias can influence complex utility calculations, making it likely that the best way to get someone to do the utility-maximizing thing is to ingrain in them a disposition to perform certain easily identifiable utilitous acts.

After having made these utilitarian arguments for following general rules, I go on in Section VI to discuss whether these reasons really help utilitarians deal with the cases that non-utilitarians argue put utilitarianism at odds with our moral intuitions. I take up the specific example of Transplant cases, and argue that utilitarians have good reason to follow the generally utilitous rule of non-interference in these cases, too.

One of the crucial points of this argument involves discussing whether one act of interfering with someone's body against their will can have significant effects on society's expectations about the future likelihood of such acts. Since this is a crucial issue—and closely linked to the one raised by Williams in the quote above—I devote all of Section VII to discussing it and arguing that, contrary to what Williams suggests, one act of interference *can* have significant negative effects on others' expectations. This is an empirical question, of course, like many of the questions I will discuss in this chapter. Empirical issues can't be avoided, however, given that the objections to utilitarianism with which we're dealing are about what sorts of actions utilitarianism actually requires, and this depends on what effects these sorts of actions actually produce. I won't be able to settle the empirical questions in anything like a conclusive way, but the goal is to make some general, often overlooked points about them.

In Section VIII, I very briefly discuss how the concerns I mentioned in relation to Transplant cases are applicable in other cases often used to object to utilitarianism. I conclude that the benefits of breaking the generally utilitous rule in these cases are not certain enough to justify a utilitarian in departing from the rule. Or where the benefits *are* certain enough, I believe our intuitions generally go along in permitting departure from the rule. Thus I conclude that utilitarianism supports employing a decision procedure with the same general mix of deontological and consequentialist principles as the decision procedure our moral intuitions prompt us to use.

I. Act utilitarianism versus rule utilitarianism

For a period of about thirty years, from roughly 1950 to 1980, considerable interest was shown in a version of utilitarianism known as “rule utilitarianism.”[191] On rule utilitarianism, it was not the consequences of individual *actions* that were to be evaluated but rather the consequences of groups of actions that all followed a particular *rule*. Rule utilitarianism said that, rather than performing that action which would bring about the best consequences if done in one particular case, one ought to perform that action which would bring about the best consequences if done in all similar cases.

There were at least three reasons that such a view was attractive. First, it seemed that morality in general was about rules and principles, and so taking rules as the point of consequentialist evaluation seemed appropriate. Second, rule utilitarianism seemed to allow utilitarians to solve certain coordination problems: it seemed to allow a group of utilitarians to produce better consequences than they would if they were all reasoning as act utilitarians. This coordination benefit turned out to be illusory, however, since acting on a principle that has the best consequences if everyone acts on it does not guarantee that others *will* act on the principle. In some

[191] The view was suggested by S. E. Toulmin, *An Examination of the Place of Reason in Ethics* (London: Cambridge University Press, 1950) and J. O. Urmson, “The Interpretation of the Moral Philosophy of J. S. Mill,” *The Philosophical Quarterly* 3, no. 10 (1953): 33-9, and versions of it were defended by Rawls, “Two Concepts of Rules,” *Philosophical Review* 64 (1955): 3-32; Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979) and *Morality, Utilitarianism, and Rights*; and John C. Harsanyi, “Rule Utilitarianism and Decision Theory,” *Erkenntnis* 11 (1977): 25-53.

cases, if one acts on the principle when others don't, one will produce very *bad* consequences. It seems, then, that one ought to act only on that principle which will bring about the best consequences *given how other people are acting*. But this is just act utilitarianism.

A third reason for interest in rule utilitarianism, however, was that rule utilitarianism seemed able to match many more of our moral intuitions than act utilitarianism. While there are many cases where special circumstances seem to make it true that transgressing our usual moral principles would have the most utility, rule utilitarianism says that we should *not* transgress our usual moral principles in such special circumstances because *on the whole* they are the principles with the best consequences, and it is at the level of rules, not of individual actions, that consequentialism is normative.

Critics of rule utilitarianism have asked, however, how one could have reason to follow a rule that is generally beneficial even in situations in which it is *not* beneficial. It seems to go against the spirit of consequentialism to reject an opportunity to improve the future, even if this means doing something which does not conform to a rule that is beneficial in other cases. Refusing to break a rule in a case in which it does not produce the best consequences seems to be, in the words of Smart, "a case of rule worship."[192]

[192] Smart, "An Outline of a System of Utilitarian Ethics," 10. Similar complaints have been voiced by Philippa Foot and D. H. Monro. See Philippa Foot, "Utilitarianism and the Virtues," *Mind* 94, no. 374

It would be an especially grave mistake for someone who endorses utilitarianism for the metaethical reasons I've given in this dissertation to adopt a traditional rule-utilitarian approach. This is because, according to the metaethic I've defended, normativity resides in the phenomenal states that result from our actions. This normativity comes to make claims on our actions just because those actions have a potential causal link to normative phenomenal experiences. Whenever an action could prevent the instantiation of a negative normative quale, for example, there is a *pro tanto* reason to perform the action. To step in and insist that such an action should not be performed *simply because it violates an otherwise beneficial rule* is to introduce an entirely irrelevant concern. On the metaethical view I've defended, a rule in itself has no normative force. When a rule should be obeyed, it is because following it will bring about certain consequences which are intrinsically normative. But if following a rule on a particular occasion will not bring about good consequences or prevent bad ones, there is no reason to follow it in that case. Furthermore, if negative normative qualia will result from following the rule on that occasion, or if positive normative qualia will result from breaking it, then there is in fact reason *not* to follow it.

Rule utilitarianism, to be metaethically justified, would have to rely on some further normative fact about the intrinsic normativity of rules as such. Having rejected the belief that any such fact exists, we must reject the idea that rules intrinsically have

(April 1985): 196-209, p. 196; and D. H. Monro, "Utilitarianism and the Individual," *Canadian Journal of Philosophy* 5, suppl. (1979): 47-62, pp. 54-5.

any privileged place in determining what we ought to do. As Smart and Lyons have argued, if the normative force of a rule is due strictly to the consequences of conformity, then rule utilitarianism collapses into act utilitarianism.[193]

Thus, if following rules other than the Principle of Utility is to be justifiable on the basis of the metaethical view I've presented—a view on which all normativity derives from the normativity inherent in the states of affairs produced—it's going to have to be because, *in the individual situation we're considering*, following these rules is the procedure most likely to lead to the best state of affairs (i.e., most likely to lead us to conform to the Principle of Utility). We can't appeal to a duty to follow generally beneficial rules in order to make utilitarianism match our intuitions. In order for it to be the case that we ought to follow the rules in a particular situation, following them has to produce the best possible consequences *in that situation*.

II. Preliminaries about the utility-maximizing decision procedure

Before we get into the details of which rules other than the Principle of Utility it could be utility-maximizing to follow, it will be useful to be a bit more precise about what I mean by "following" rules. In more precise terms, the central claim of this chapter is that *employing a decision procedure that includes rules other than the*

[193] See Smart, 9-12, and Lyons, *The Forms and Limits of Utilitarianism* (Oxford: Clarendon Press, 1965). For similar considerations, see Hare, *Freedom and Reason* (Oxford: Clarendon Press, 1963), 131-6; and Brandt, "Toward a Credible Form of Utilitarianism," in Hector-Neri Castañeda and George Nakhnikian, eds., *Morality and the Language of Conduct* (Detroit: Wayne State University Press, 1963), 107-43, pp. 119-23.

Principle of Utility can lead us to conform (as closely as is possible) to the Principle of Utility. Section I showed that the metaethical view defended in Chapters 3 through 5 cannot justify employing a decision procedure which leads us to violate the Principle of Utility. If it is going to justify our making decisions according to rules other than the Principle of Utility, this will have to be because employing those rules as a decision procedure leads us in the end to act in such a way as to conform as closely as possible to the Principle of Utility.

It should be obvious that conforming to the Principle of Utility requires employing a decision procedure other than the conscious calculation of all the expected utilities of all one's possible actions. Calculating utilities requires a great deal of information, time, and energy. To do it perfectly, one would have to have complete knowledge of the present state of the universe and of all the future states that could possibly follow from any of the different actions one could perform. Certainly no human being could ever have all of this information, and it may well be impossible for any agent ever to have it all. Furthermore, even if one could have all this information, collecting it and making calculations based on it would waste time and energy that would produce more utility employed in some other activity. And even if one were simply to rely on the information one already had about expected utilities, the simple act of mentally running through all of the possible consequences of an action and its alternatives would be too time- and energy-consuming to maximize utility, except perhaps on the occasion of a particularly momentous decision. For these reasons, it should be clear that one cannot conform to the Principle of Utility by regularly employing the Straightforward Utilitarian Decision Procedure.

What we need is to employ a decision procedure other than the Straightforward Utilitarian Decision Procedure (hereafter, the "Straightforward UDP") which nevertheless leads us to perform the act with the best consequences. One might wonder how any decision procedure besides the Straightforward UDP could reliably lead us to perform the act with the best consequences. In his discussion of rule utilitarianism, Smart argues that "any rule which can be formulated must be able to deal with an indefinite number of unforeseen types of contingency. No rule, short of the act-utilitarian one, can therefore be safely regarded as extensionally equivalent to the act-utilitarian principle unless it is that very principle itself."[194] But while Smart is right that no rule besides the Principle of Utility will identify the optimal act *in all circumstances*, it is nevertheless true that another rule may identify the same acts as the Principle of Utility *within a particular context*. If we know ahead of time certain features of the circumstances that an agent will confront, we may be able to give him a decision procedure that is easier to employ than the Straightforward UDP and yet leads him to choose the same acts.

Consider the extreme case in which we know *exactly* what situation a particular agent will confront. We know everything about the situation and everything about all the consequences of any action he might choose. Then we know which action

[194] Smart, 12.

he ought to perform according to the Principle of Utility. Instead of telling him, "Follow the Principle of Utility," we can tell him, "Do *X*." And by doing *X* in that situation, he will actually conform to the Principle of Utility.

The crucial points are these:

(1) The Principle of Utility does not require us to employ the Straightforward UDP.

(2) Employing the Straightforward UDP wastes time and energy, thus actually causing us to violate the Principle of Utility.

(3) There may be a decision procedure tailored to the demands of actual situations in such a way that, by employing it, we end up conforming much more closely to the Principle of Utility.

Determining the broad outline of this decision procedure, and arguing that it generally conforms to our moral intuitions, is the goal of the rest of this chapter.

The first step is to investigate what sort of decision procedure would be most likely to maximize utility. We know that no decision procedure is as adaptable to all sorts of contingency as is the Straightforward UDP, but all we need is a decision procedure adapted to the features of actual situations. In the next three sections, I'm going to discuss three general features of the situations human beings actually face and what implications they have for the sort of decision procedure it is utility-maximizing for us to employ. These three features are uncertainty, the need for coordination, and motivational limitations.

Of course, there’s no way that I can prove that a particular decision procedure will be utility-maximizing in every situation we’ll face. This would require as much time and information as employing the Straightforward UDP in all of these cases. At best, I can point to some general evidence that certain systematic features of the situations in which we find ourselves make this likely to be the case.

On the other hand, I want to make it clear that at no point do I intend to rely on the claim that we ought to follow a rule in *all* situations just because following it is utility-maximizing in *most* of them. I am simply arguing that, in situations that have certain features, we do best to employ a decision procedure which contains certain rules. If we have sufficient reason to believe that a particular situation does *not* have certain of these features, however, this will be sufficient reason to deviate from this decision procedure. And in such cases, I believe deviation from this decision procedure will not be contrary to our moral intuitions.

III. First general feature of actual situations: Uncertainty

One of the most important features of the situations in which we actually find ourselves is uncertainty. We simply never have enough information to be absolutely certain of the consequences of our actions. Thus, in order for us to be able to employ a decision procedure, it obviously cannot require us to make decisions based on the actual consequences of possible actions; it can only require us to make decisions based on predictions about their consequences.

To make these predictions, we are forced to rely on information about past consequences of relevantly similar actions. We have to rely on generalizations of the form: Actions of Type *A* in situations of Type *B* on average have utility *x*. But to know which generalization to rely on in a particular case, we have to know how broadly or narrowly to characterize the type of action and type of situation. We have to determine which past act-situation pairs give us the most accurate information about the one we're now considering.

Consider a woman who is contemplating marrying a certain man and wants to know what the chances are that such a marriage will make her happy. She has to determine which past marriages are relevantly similar to this prospective one to give her useful information about her own chances at happiness. Does she take as her point of comparison the average outcome of all marriages? All marriages in the past century? All marriages in her country in the past decade? All marriages in her family and the man's family for the past two generations? All marriages where the age difference between the partners is similar to the difference in the present case? All marriages between partners of social classes like theirs? Perhaps she shouldn't take into account only marriages but also all long-term relationships. Perhaps if she plans to have children, she should only take into account those marriages where the wife desired children. Or if she plans to have a professional career, perhaps she should only take into account those marriages in which the wife had a profession similar to hers.

Fortunately, we do not have to establish any level of generality in the description of an act-situation pair as distinctively normative. We need only aim for the level of generality which is most likely to give us the most accurate prediction, balanced against any costs of making the prediction more accurate. The two main points to consider in aiming for accuracy are (1) that the more similar all the comparison cases are to the present case, the more likely they will be to yield an accurate prediction, and (2) that the more similar one requires the comparison cases to be, the smaller the overall sample will be, and the easier it will be for some random factor to distort the prediction. Thus one should not go to either extreme: one should not overly restrict one's sample size, but neither should one include cases which are so different from the actual case as to be irrelevant. Sample size and relevance have to be balanced against one another, in view of producing the most accurate prediction possible. Of course, in order to know just how to balance them, one must rely on statistics about the success of past predictions based on various sample sizes, and the generality problem starts all over again when one tries to determine which past predictions are enough like the present one that their sample sizes are relevant!

But in fact recent research indicates that the situation is not nearly as complicated as one might expect. In his 2007 book *Gut Feelings: The Intelligence of the Unconscious*, Gerd Gigerenzer assembles research showing that the most accurate way to predict complex phenomena is very often to focus on the one factor that is the most closely correlated with the effect in question. This means that, if research shows that the single best predictor of happiness in marriage is one's average happiness with a romantic relationship in its second year, our agent will make the most accurate prediction of her happiness being married to *this* individual by relying solely on this factor.

Now common sense normally tells us that, if we have the time and energy to consider more factors, we will get a more accurate prediction by doing so. But this is precisely what the research assembled by Gigerenzer contradicts. Considering each additional factor after the first severely reduces one's sample size, and random effects are more likely to distort one's prediction than if one retained the larger sample size. Granted, in some cases, the amount of data available to an agent may be so enormous that random effects will still cancel each another out even if the agent considers two or three factors, but on the whole, considering one strongly influential factor will produce the most accurate predictions. Gigerenzer summarizes the research thus:

> Intuitions based on only one good reason tend to be accurate when one has to predict the future (or some unknown present state of affairs), when the future is difficult to foresee, and when one has only limited information. They are also more efficient in using time and information. Complex analysis, by contrast, pays when one has to explain the past, when the future is highly predictable, or when there are large amounts of information.[195]

[195] Gerd Gigerenzer, *Gut Feelings: The Intelligence of the Unconscious* (New York: Viking, 2007), 151. This paragraph summarizes results reported in the following six works: G. Gigerenzer, P. M. Todd, and the ABC Research Group, *Simple Heuristics That Make Us Smart* (New York: Oxford University Press, 1999); K. Katsikopoulos and L. Martignon, "Naïve heuristics for paired comparisons: Some results on their relative accuracy," *Journal of Mathematical Psychology* 50 (2006): 488-94; L. Martignon and U. Hoffrage, "Fast, frugal and fit: Lexicographic heuristics for paired comparison," *Theory and Decision* 52 (2002): 29-71; R. M. Hogarth and N. Karelaia, "Simple models for multi-attribute choice with many alternatives: When it does and does not pay to face tradeoffs with binary attributes," *Management*

Thus we conclude that the decision procedure most likely to optimize utility in actual situations will tell us to predict utilities based on only the one factor best correlated with the effect in question (or perhaps on the two or three best correlated factors, where the sample size is particularly large). The high level of uncertainty present in actual situations makes it the case that we have the greatest chance of bringing about the best consequences if we make decisions according to *very general* rules, such as "If you're very happy in your second year of dating someone, marry them. If you're not, don't."[196]

IV. Second general feature of actual situations: Need for coordination

Yet not only is making decisions according to very general rules the best way to deal with uncertainty about the consequences of our actions, it can also *reduce* this uncertainty. Much of the uncertainty about our actions' consequences results from the fact that their consequences depend on the actions of others. We often can't know which action of ours will be utility-maximizing if we are ignorant of the decisions that

Science 51 (2005): 1860-72; R. M. Hogarth and N. Karelaia, "Ignoring information in binary choice with continuous variables: When is less 'more'?" *Journal of Mathematical Psychology* 49 (2005): 115-24; and R. M. Hogarth and N. Karelaia, "Regions of rationality: Maps for bounded agents," *Decision Analysis* 3 (2006): 124-44.

[196] Note, however, that a rule based on past experience could be unreliable in a case that is missing some feature common to all of the cases from which one's data was collected. This feature may or may not be important, but data limited to cases where it was present cannot tell us. If, for instance, all of our data about happiness in marriage is taken from Western, individualistic societies, we may come up with a rule that says one ought not to marry someone before dating them for at least eighteen months, but in other societies, where family and community play a larger role in marriage, this delay may be unnecessary. Only data including these other types of society can tell us.

others will make. For instance, it might be the case that, if another person were to choose A over B, the best overall consequences would result if we chose X rather than Y. But it might also be the case that, if the other person were to choose B, the best overall consequences would result if we chose Y. When we don't know whether the other person will choose A or B, we don't know whether to choose X or Y.

A practice of making decisions according to general rules could resolve this dilemma. If we knew that other people generally employed a certain rule when choosing between A and B, then we would have a basis for knowing which they would choose and thus a basis for deciding ourselves whether to choose X or Y.

Now perhaps it will be suggested that, if everyone was a utilitarian, we could always rely on others' making the choice they believed would have the best overall consequences. We could predict their behavior by their utilitarianism, without needing them to employ any more specialized rule. But in fact what we need in order to be able to predict the behavior of others is not just that they employ *some* rule, such as the Principle of Utility, but that they employ a rule whose application to any particular case is obvious enough that we can easily predict how they will apply it. We don't have the time or energy to consider, in addition to all of the factors that have a direct influence on the outcome of our own actions, all of the factors that influence the utility calculations of others (including their beliefs about what decisions *we* will make). What we need is for others to make decisions, not by calculating all of the probable consequences of their actions, but by employing some simpler rule, at least in those cases in which their actions directly influence the consequences of ours.

A prime example of such a useful, simpler rule is the rule of promise-keeping. A practice of promise-keeping allows us to establish with someone ahead of time the manner in which they will behave and allows us to rely on that behavior in making our own utility calculations. If someone promises me that she will go to dinner with me on a certain night and I believe that she will keep her promise, then I will base my further plans for the day around this expectation. Perhaps I will eat less at lunch in anticipation of a large meal at night, or spend more time working during the day so that I won't have work left to do in the evening. I may spend time during the day thinking about the meal and taking pleasure in anticipating it. I would not do these things if I wasn't sure that my friend would really go to dinner with me, because in that event, these same actions could make me worse off. It might have been better for me to eat more at lunch so that I had more energy for the rest of the day, and better for me not to anticipate the dinner all day long because of the strong disappointment I feel when I discover it's not going to happen. What it's best for us to do depends in large part on what others do, and when others' actions have a particularly strong influence on the results of our ours, we do well to coordinate our actions with theirs. One way in which we can do this is through making and keeping promises.

Now it's not the case that we will always maximize utility by keeping a promise, nor that we will always be most *likely* to maximize it in doing so. Sometimes it’s very clear—as when a child will likely drown unless we miss our dinner date—that we ought to break a promise. Whether we ought to break a promise on any particular occasion depends on several factors. One of the most important is how great the benefits of breaking the promise are likely to be. But this we will have to measure against the probable *dis*utility of breaking it. The probable disutility is going to depend on what sorts of things the person to whom we’ve promised is likely to have staked on our keeping the promise. This will depend on what their expectations are for our keeping it, and that in turn will depend on their knowledge of us personally, as well as on what the general cultural expectations are about keeping promises of this kind. People tend to stake relatively little on promises of dinner dates. They stake a great deal more on marriage vows. But they will stake more or less on each of these kinds of promises depending on their beliefs about the person making them. If someone stands them up once or twice, they stop staking anything on that person showing up in the future.

Does this mean that if someone doesn’t expect us to keep a promise, we don’t have any reason to keep it? Not necessarily. There is yet a further reason to keep promises: to increase expectations of their being kept in the future. As we’ve already noted, it’s useful for us to be able to coordinate our actions with one another, to be able to indicate to one another how we will be acting in the future so that others can make more utilitous decisions. In order for the practice of promising to work, however, the act of saying, “I promise to do *X*.” has actually to raise the promisee’s expectation that the speaker will do *X*. Once their expectation is raised, the promisee will make certain decisions counting on the promisor's doing *X*, and the promisor's doing *X* will then be more utilitous than his not doing it, *ceteris paribus*.

There may be some difficulty in raising this expectation, however, if one is known to make one's decisions by calculating the utility of their consequences. D. H. Hodgson has argued that, in a society of persons all known by each other to be attempting with high rationality to act according to the Principle of Utility, there will be no way for the expectation of promise-keeping to get started. He writes,

> …a promised act could have greater (comparative) utility (than it would have had if it had not been promised) only if the promisee has a greater expectation that it would be done (than he would have had if it had not been promised); but there would be a good reason for such greater expectation only if (in the promisor's belief) the act would have such greater utility. Being highly rational, the promisor would know that the greater expectation was a condition precedent for the greater utility; and so would not believe that the act would have greater utility unless he believed that the promisee had greater expectation. Also being highly rational, the promisee would know this, and so would not have greater expectation unless he believed that the promisor believed that he had greater expectation.[197]

Now if the promisee could simply let the promisor know that she would act relying on the promisor to do *X*, the promisor would see that he had act-utilitarian reason to do *X*. Similarly, if the promisor could simply wait and see whether the promisee acted out of reliance on his promise before he had to decide whether to keep the promise or not, the promisee would have act-utilitarian reason to rely on the

[197] D. H. Hodgson, *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory* (Oxford: Clarendon Press, 1967), 41.

promise’s being kept. This is because, once the promisee has acted in reliance on it, the consequences will be better if the promise is kept, and the promisee knows that the promisor will know all of this and act to produce the best consequences. But the practice of promising is supposed to enable us to tell people what we *will* do, so they can count on its happening before we’ve actually done it, or when they’re unable to find out whether we’ve done it. And this is something that two people known to each other always to follow the Straightforward UDP cannot communicate between themselves, for the following reason:

If A and B are two agents each known to the other always to follow the Straightforward UDP, A will only count on B’s doing X if A knows B knows A is counting on B’s doing X. But B will only know A is counting on B’s doing X if B knows A knows B knows A is counting on B’s doing X. And B will only know this if he has some further knowledge about A’s knowledge, which he will only have if he has even further knowledge about A’s knowledge, etc. If both parties are known to each other always to follow the Straightforward UDP (and known to be known always to follow it, and so on, as in Hodgson’s example), then they will each need an infinite amount of knowledge about each other’s knowledge in order to make the coordination happen.

A promise is designed precisely to stop this informational demand, by making it known that one agent is going to act in a certain way independently of his beliefs about the other agent’s behavior. But as Hodgson points out, if B always follows the

Straightforward UDP, he will *not* act independently of his beliefs about A. And if A knows that B always follows the Straightforward UDP, he knows that, even if B says "I'm going to do X," B will only do it if he knows that A knows that B knows that A knows that B knows…*ad infinitum*…that A is counting on B's doing X. The problem is that making a promise or stating an intent will not change others' expectations about our behavior unless it is believed to reflect an actual disposition of ours to do the act in question. But in the case where the consequences of the acts of two agents known to each other always to follow the Straightforward UDP are dependent on one another, a promise or statement of intent can never reflect a disposition to perform the act in question because the promisee never has enough information to justify his performing it on the basis of the Straightforward UDP.

We can avoid this sort of coordination problem, however, if people have even a very slight disposition to deviate from the Straightforward UDP. What is needed is a disposition that will allow other people, when they are unsure of our utility calculations, to trust rightly that we are more likely to keep our promises than not, and this can be accomplished be a very minimal departure from the Straightforward UDP. All that's necessary is that, *in cases where we are unsure of the utility of keeping a promise, we be disposed to err on the side of keeping it.*[198] Once others believe we are

[198] Of course, there's the possibility that one could solve the coordination problem by *deceiving* others into thinking one has certain dispositions when in fact one does not. However, it's hard to see how this could be the more utilitous approach. Once others' expectations are raised, actually having the dispositions is just as utilitous as continuing to make decisions strictly in accordance with the Straightforward UDP, and the non-deceptive path doesn't require expending the effort to deceive. It

more likely to keep our promises than not, they'll base their own utility calculations on this belief and we will then have a straightforwardly utilitarian reason to keep them.

Luckily, most of us already have a disposition to keep our promises even in cases where we're uncertain of the utility of doing so.[199] And in fact we also have other dispositions that depart from the Straightforward UDP but that are useful for promoting coordination. These are dispositions to conform to societal norms, at least where there is no clear utilitarian reason to depart from them. Since we can't make promises about every one of our actions, it's very utilitous for others to know how we are generally likely to behave, from the fact that we are more likely to tell the truth than not, to the fact that we will tend to follow accepted ways of doing business, engaging in small talk, managing a university classroom, dealing with colleagues, dating and marrying, and all manner of other things. And of course once others are expecting us to behave in a certain way, we have a straightforward utilitarian reason to do so (although one that may be outweighed by other reasons on any particular occasion).

While the departures from the Straightforward UDP needed to get coordination going are rather minimal (only coming into play when we are on the fence about expected utility), we should note that they are nevertheless of a somewhat stronger

seems that the most utilitous way to solve utilitarian coordination problems will be actually to have the dispositions it's utilitous for others to expect us to have.

[199] This is lucky because it means that we don't have to worry about the feasibility of converting ourselves from being strict followers of the Straightforward UDP to having such a disposition. Although I think this could be done (by manipulating the motivations of our future selves), it's a complicated issue and fortunately not one that has to be gotten into here.

nature than those necessary to solve the problems of uncertainty described in the last section. In the last section, we saw that uncertainty in predicting the future necessitated the use of what are often called “rules of thumb.” Following these rules of thumb, however, is not strictly departing from the Straightforward UDP. In following them, we are still basing our decisions purely on utility calculations; it’s just that these calculations have to be very general because of our lack of information. To solve coordination problems in the most utilitous way, however, we have to be disposed in certain situations to base our decisions on something other than utility calculations. In situations of great uncertainty, at least, we need to be disposed to keep promises and to do other things that others will expect, even if we have no reason to think this is more utilitous than the alternative.

This is very similar to the conclusion that John Gray draws from utilitarian coordination problems. He writes,

> If, as Hodgson and the indirect view both maintain, direct utilitarian policy erodes the practices necessary to social cooperation, then these practices must be supported on utilitarian grounds as *imposing constraints on utilitarian policy*. They cannot be merely the rules of thumb of which Smart speaks: rather, insofar as they do constrain utilitarian action and deliberation, they possess what might be called ‘second-order’ utility of their own, which they must lose if they are to be regarded as always vulnerable to utilitarian overriding.[200]

[200] John Gray, “Indirect Utility and Fundamental Rights,” *Social Philosophy and Policy* Vol. I, Issue 2, Human Rights, edited by Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul (Oxford: Basil Blackwell, 1984), 73-91, pp. 83-4.

Gray is right that the solution to utilitarian coordination problems demands more than following rules of thumb. He is also right that what is needed is some non-utilitarian reasoning. What he misses is that, while the reasoning that is needed is *non*-utilitarian, it is nevertheless of such a kind that it does not actually *conflict* with utilitarian reasoning. It is invulnerable to utilitarian overriding simply because it only comes into play when utilitarian reasoning is inconclusive.

To solve utilitarian coordination problems, all we need is a disposition to keep promises and conform to others' expectations in those cases where we're uncertain about the utility of doing so. This is enough to ground others' expectations, and once these expectations exist, we will have a more direct utilitarian reason to conform to them: the fact that others' utility calculations have been based on them.[201]

V. Third general feature of actual situations: Motivational limitations

We turn now to a third problem with employing the Straightforward UDP: having the motivation to do so. What we are looking for is the best possible decision procedure for actual human beings to employ in actual situations, and this means that we may have to reject certain decision procedures if they are impossible for human

[201] One might wonder whether this explanation can justify keeping promises made to someone on their deathbed. It can if we believe that the person asked us to promise to do the thing in question because of the positive consequences of doing it, which may be known only to him. If, however, we can be certain that keeping a promise to someone dead will have negative consequences overall, utilitarianism tells us that we shouldn't keep it, just as we shouldn't keep promises to living persons in such a situation.

beings to employ—or at least impossible for them to employ entirely correctly—due to certain limitations on their motivational systems.

Probably the clearest of our motivational limitations is a bias towards our ourselves and those we love. It seems likely that it is not possible for human beings to be motivated in a straightforwardly utilitarian way—to be equally concerned about the interests of all experiencing subjects. Some preference for ourselves and for those we love is probably ineradicable. But if we do have these motivations that are impossible to change (or that are at least prohibitively costly to change), then the optimal decision procedure is going to be one that somehow works *with*, rather than against, these motivations.

One thing we can do to greatly increase utility-maximization given such limitations is to create a strong causal connection between each person's actions and the positive consequences they would be most motivated to bring about. Because people are highly motivated to perform actions they think will have positive effects on their own happiness or the happiness of those they love, and because they are much less motivated to perform actions whose benefit is uncertain or goes to someone they don't know, a utility-maximizing society will be ordered in such a way that each person's actions have a strong link to their own happiness and that of their loved ones.

Perhaps it might seem that each person's actions *always* have a strong effect on their own happiness and that of those they care about. Consider, however, that such an effect would not be assured in a society in which zealous utilitarians were always interfering in others' lives in an attempt to maximize total utility. (Consider the interference that goes on inside a totalitarian regime, even one with laudable ultimate goals.) If we didn't generally recognize a duty to refrain from interfering in others' projects, if we didn't allow that others could have property which—being theirs—is off-limits to us, then each of us would find our lives a chaotic mess of interference from others. We would have very limited possibilities for promoting our happiness because of the way any investment we made in our future could be taken away from us without warning. In such a world, no one would be motivated to undertake any projects other than those producing instant gratification, because they would almost certainly be doomed to failure. On the other hand, by generally being disposed to respect other's "rights"—rights to certain degrees of non-interference with their lives, limbs, and property—we foster the kind of environment in which each person is motivated to work and invest in their future happiness.

This sort of motivation would also disappear in a society in which, though there was little interference in the operation of each agent's projects, the benefits of their projects were not strongly felt by the agents themselves but were always spread out across a large number of strangers. Human beings simply don't seem wired so as to be able to sustain high motivation to perform actions that return no sizable benefits to them or their loved ones.

Of course, the considerations I've given so far don't make clear exactly what each person's sphere of freedom from interference ought to be. It seems clear that it's important for one's sphere of sovereignty to include one's body. However, what amount or degree of personal property is necessary to produce optimal levels of motivation is not as clear. Perhaps a system of semi-communal property would have better consequences overall than a system of wholly personal property. Couples seem capable of sharing a large amount of their property without a significant decrease in motivation. The success of certain communes has shown that even larger groups can share a great deal of property while maintaining individual motivation to use it relatively profitably. To determine what precise system of property is optimal, we are going to need the help of more empirical research. It's clear, however, that we're going to need to be able to expect the use of *some* material objects without threat of interference from the majority of humankind.

In many situations, it's also useful to have certain spheres of freedom from interference that are defined not by material objects but in more abstract ways. In cooperative endeavors, especially more complicated ones like managing a government or a business, it's useful to be able to designate individuals to have charge of particular areas of the enterprise which are better managed when everyone is not trying to interfere with their operation. Generally, decisions about which people ought to be in charge of which areas are very openly made on the basis of utility considerations. And once these areas have been designated, encroachment on them is often seen as a violation of the individual's "right" to make decisions in that area, a right which there is generally utilitarian reason to respect, since that individual likely has better than average knowledge of her area, some cohesive strategy for dealing with its complex problems that will be upset if there is outside interference, and the motivation to see that things in this area go well, since negative outcomes will reflect directly on her management.

So the exact boundaries of the spheres of sovereignty we ought to respect will depend on a number of factors affecting which property system and which division of responsibility are likely to be utility-maximizing. Of course, in making actual decisions about non-interference, we need to consider not just what sphere of sovereignty an ideal system would grant each person, but also what people's existing expectations for non-interference are, since they are likely already to have based many utility calculations on these expectations, and since any process of expectation-change will likely have to be somewhat gradual.

I now want to turn to a second problem caused by self-interested bias, and this is that we are not as capable of purely rational reflection as we might like to think. However objective we think we're being when we're determining the harms and benefits of a particular action, our deeply ingrained concern for our own interests and for the interests of those we love can seep into our judgments without our knowledge. While we may think we are making an objective assessment of the consequences our action will have on some faraway strangers, we are in fact quite likely to underestimate the negative effects our actions will have on others when the benefit to ourselves is great. We are excellent at unconsciously rationalizing our self-interested behavior by constructing what seems—even to us—like an objective argument in support of it. And the more complex and uncertain the issues we're reflecting on, the greater the opportunity our selfish bias has to slip into the calculations undetected.

This suggests, once again, that the best way to get ourselves to do the utility-maximizing thing is not to perform a complex calculation of the relevant utilities. In view of the ability of our self-interested bias to work itself into our calculations without our knowledge, a more effective alternative to calculating utilities in individual cases might be to determine ahead of time (i.e., before we're in a situation where we have a particular interest) which kinds of actions, classified according to easily identifiable characteristics, have the highest and lowest average utilities and to instill in ourselves a strong disposition to do the former sorts of actions and refrain from the latter.

We normally think of this sort of dispositional training taking place during childhood, when parents reward politeness, generosity, compassion, honesty, fidelity, and respect for others' bodies and property, and discourage avarice, boastfulness, selfishness, deception, and violation of others' rights. New behavioral dispositions can also be acquired in adulthood, however. The expression of approval or disapproval by a large number of our peers—or by a small number of people whose opinions are especially important to us—is quite effective at changing our behavior, and we should use this influence to encourage others to shape their dispositions in utilitous ways, as well as to encourage them to encourage *us* to do the same.

What’s crucial about this encouragement is that it be directed toward actions that are very easily identifiable. It’s important that society express disapproval of the act of lying in general rather than merely the act of lying in those cases where the utility of telling the truth is greater. If, to determine whether he ought to do something, an agent has to reason about whether his act really qualifies as one of the acts disapproved of, his self-interest has the opportunity to creep in and convince him that the negative consequences for others of that particular lie are not as great as the positive consequences for himself. But, if our dispositions are such that some very clear characteristic of our action or situation motivates us to act without going through any complicated reasoning process, we bypass the opportunity for our self-interested bias to distort our reasoning.

It might be thought that the fact that our self-interest is such that it has to be countered with such general dispositions is unfortunate. We might think it’s too bad that we have to trade off more refined utility discriminations for increased motivation. But, as we saw in the discussion of uncertainty, we actually increase the expected utility outcome of our decisions when we base them on fewer variables. We have a purely statistical reason to want our decisions to be based on very general features of our situation, and our need to avoid complex utility calculations in order to avoid self-interested bias simply gives us another reason to desire this. So in fact there’s no trade-off: the reasons in both areas are in favor of our being disposed to perform and avoid some *very* general sorts of actions.

If we combine the conclusions of this section with those of the last two, we see that there are in fact three overarching reasons that it's utility-maximizing to be disposed to perform and avoid some very general sorts of acts. The first reason is that we have the greatest chance of performing the most utilitous act if we select it on the basis of only one or a very few factors: those factors which are the most strongly correlated with positive utility. The second reason is that others' utility calculations are made more accurate if they can predict how we will act, and this they will only be able to do if we make choices based on a minimal number of factors, and specifically, factors of which they are also likely to be aware. The third reason is that decisions based on more complex calculations are more likely to be biased by self-interest. From all of this, we conclude that the decision procedure most likely to maximize utility in a situation which contains uncertainty, the need for cooperation, and unconscious motivational bias is a decision procedure which prescribes and prohibits certain broad *categories* of action, without discriminating between more and less utilitous actions within these categories.

Normally, if a moral theory categorically prescribes or prohibits actions in this way, we call it "deontological," and contrast it with "consequentialist" theories. But we've just shown that, in situations with certain epistemic and motivational limitations, consequentialism actually requires agents to adopt a decision procedure with a deontological form. It requires agents to employ general rules, not because rules of this level of generality are particularly normative, but because, *in any individual* *situation where there is uncertainty, a need for coordination, and unconscious motivational bias*, following a very general rule is most likely to maximize utility.

The last three sections have also given us an idea of the content of the very general rules we ought to follow. Among other things, we will normally have the best chance of maximizing utility if we keep promises, tell the truth, and refrain from interfering with others' bodies, property, or areas of responsibility. The optimal decision procedure will of course allow exceptions to these rules in those cases where it's *extremely clear* that it's utility-maximizing to break them, since these are cases in which concerns about uncertainty and unconscious bias do not apply. And for decisions that don't involve promise-breaking, lying, or interfering with others' bodies, property, or areas of responsibility, the optimal decision procedure may have one follow a more self-consciously consequentialist approach (though it may permit one to focus most of one's attention on problems close to home, if these are the ones one is likely to be most effective at solving[202]). This sort of decision procedure—one which prohibits certain actions unless utility is overwhelmingly on the side of performing them (and perhaps allows certain other actions unless utility is overwhelmingly on the side of not performing them)—looks an awful lot like the decision procedure our moral intuitions prompt us to follow.

[202] For a very brief discussion of the utility of agent-relative permissions, see Section I.4 of Chapter 5.

VI. Applying this decision procedure to Transplant cases

The arguments of the past three sections have been intended to show that the general shape of our moral intuitions—their inclusion of certain deontological constraints (or permissions) within a larger consequentialist framework—matches the general shape of the decision procedure that is most likely to maximize utility in actual situations. These arguments don't prove, however, that there are no actual situations in which utilitarianism requires us to do something at odds with *our precise level* of intuitive respect for keeping promises, telling the truth, and refraining from interfering with others' bodies, property, and responsibilities. Whether the utility-maximizing decision procedure tells us to stick to these rules in exactly the same situations as our intuitions depends on (1) the situations in which our intuitions tell us to stick to these rules, which will vary somewhat from person to person, and (2) the situations in which we are capable of accurately judging that breaking the rules will be of greater utility than following them.

To get the most accurate comparison of the lines drawn by our intuitions and by the utility-maximizing decision procedure, we will need a lot of empirical data about just when it is utilitous to break generally useful rules, and how good we are at determining when this is the case. We should certainly try to collect more of this data, as it will put our ethical debates on much firmer ground. But in the meantime, there may still be some helpful things to say about the question. We can look at some of the individual situations in which opponents of utilitarianism have claimed it requires acting contrary to our intuitions, and we can try to estimate, for these particular cases, the strength of the reasons I've given for following general rules.

I propose that in this section we look at what I'm going to call "Transplant" cases: cases in which we know with reasonable certainty that five hospital patients will all soon die if they don't receive various organ transplants. It's often been thought that, in these cases, conforming to the Principle of Utility requires us to kill one healthy person and distribute his organs among the five sick patients, since sparing five lives and losing one would be better than sparing one life and losing five. Our moral intuitions, on the other hand, tell us that we ought not to do this. Here I want to discuss at length the various factors that influence the utility of interfering with someone's body by harvesting their organs in this way, under these circumstances. My aim is simply to show that the utility calculations are complex enough that it's not at all certain that the benefits of interference in this sort of case outweigh the costs. This uncertainty makes this just another case in which we have reason to follow a general rule, a rule whose utility has been tested across enough cases that random effects have been canceled out.

I have broken down into four main groups the factors that influence the utility of killing the one to save the five in Transplant cases. I will address them in the following order: (1) basic medical concerns, (2) the neglect of superior alternatives, (3) the drawbacks of secrecy, and (4) the consequences of discovery.

1. *Basic medical concerns*: One of the most obvious concerns we ought to have if we are actually contemplating performing these transplants is just how great the benefit would be to those who would receive the organs. The benefit to the five will clearly be less than five times the benefit the one would receive from staying alive.[203] Transplants do not have a one hundred percent success rate, first of all. There is the chance that one or more of the organs will be rejected. Second of all, to prevent rejection, the recipients will have to take immuno-suppressant drugs for the rest of their lives, drugs which can have serious negative side effects. Furthermore, the bodies of the recipients have likely already been put under great strain, either as a result of their organ failure or as a result of associated health problems. In all, we should not expect the organ recipients, on average, to have the same quality and length of life as the healthy person whose organs we harvest. The cost-benefit analysis here is not 1 to 5, especially when we factor in all of the resources necessary to perform these transplants—resources which would likely do a great deal more good if directed elsewhere.

2. *The neglect of superior alternatives*: It's very important that we remember that the mere fact that an action would produce more pleasure than pain does not mean

[203] Of course, this may not be true in every case. There may be other differences between the five and the one that make the potential benefit to the five greater and the potential benefit to the one smaller. Instead of dealing separately with all such possibilities, however, I'm just going to discuss the "average" case, where other differences among the six persons average out. The conclusion of this discussion will tell us what we ought to think about the utility of such transplants *in general*. This does not mean that an exceptional case can't arise in which the utilities are markedly different, only that, if we have no reason to believe that a case deviates much from the average case, we ought to believe that the expected utility of the transplants will be close to what is discussed here.

that the Principle of Utility sanctions it. The Principle of Utility requires one to perform the action with the *highest* possible balance of pleasure over pain. In addition to the fact that the resources needed to perform the transplants might produce more utility elsewhere, we must remember that we cannot justify killing one to save five if there is some other way that we could have saved the five *without* killing the one.

Consider that an alternative way to save five lives is to wait for the first of the sick patients to die and to use his organs to save the other four sick patients. Five lives are saved without risking all of the potentially negative consequences that would result from violating someone's right not to be killed.

Even if for some reason it would not be a good idea to acquire the necessary organs in this way, there are many other ways in which one might acquire them without violating anyone's right not to be killed. One might convince someone dying of some other ailment or injury to become an organ donor. One might lead a campaign to sign up more organ donors, so that the next person killed in a car crash will be more likely to be an organ donor. Or one might take the organs of a dead person who hadn't volunteered them. While this last option is still probably not the best, stealing the organs of a dead person is certainly preferable to stealing the organs of someone living!

It seems very unlikely that harvesting the organs of a healthy person will be the best possible way to save five lives. But furthermore, even if in some isolated case this is the best solution, it's important that we be very reluctant to pursue this course of action. We need to feel a very great repugnance for killing one to save five because this repugnance will force us to discover better options in those cases where they exist. Perhaps only an absolute refusal to kill one to save five will lead us to find the better solutions in all the cases where they are available. What is clear is that we need a strong repugnance for such killing if we aren't to do it where better options exist.

3. *The drawbacks of secrecy*: It is usually assumed, in discussions of Transplant cases, that the fact that one has harvested the organs of a healthy person to save five others will have to be kept a secret, in order to prevent various negative effects: causing fear in others, getting the person who harvests the organs sent to jail by a state that shouldn't leave such an act unpunished whether or not the act was ultimately utilitous in that particular case, and reducing people's expectations of non-interference in their lives. I will discuss the consequences of a lack of secrecy in a moment; first I want to discuss the disadvantages of secrecy itself.

We have very general reasons to avoid acting secretly, in all sorts of situations. The most important of these is that acting secretly cuts us off from the counsel and aid of others. We've already discussed our significant epistemic and cognitive limitations. Acting in secret only increases them. When making a decision about what to do in a difficult, unusual case, it would be a particularly serious mistake to rely only on one's own information and one's own reasoning ability. We need all the help we can get. By deciding early on in the decision process that we ought to keep our decision a secret, we are apt to miss some crucial information. We simply should not be confident in our assessment of the utilities involved if we have not discussed the problem with others. And the more unusual the problem, the more counsel we ought to seek out. This certainly applies to a decision as complicated as harvesting one person's organs to save five others. There are too many variables involved for us to be sure that we can cover all of the bases ourselves. The probability of making serious errors is very high if we choose to deliberate and act in secret with regard to such a complex matter.[204]

A further problem with acting in secret—even if one is not found out—is that it increases the uncertainty of others' utility calculations. To hide what we've done, we either make certain information about our situation unavailable to others or we end up having to give false information. Both of these possibilities are likely to have overall negative effects. They may be relatively minor, but it is difficult to be sure of this, especially if we are not able honestly to discuss the situation with others.

4. *The consequences of discovery*: But of course a primary danger in performing an act that is best kept secret is the possibility of being found out. And a further problem is that it is difficult to determine the exact probability of being found out. However, the more complicated the act that we have to carry out, and the more closely our actions are monitored, the more likely it is that we will be discovered. Taking organs from a healthy individual in the highly monitored environment of a hospital seems like an act that will be particularly difficult to keep under cover. We

[204] For more on the utility of honesty and transparency, see Allan Gibbard, "Utilitarianism and Human Rights," *Social Philosophy and Policy* Vol. I, Issue 2, Human Rights, edited by Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul (Oxford: Basil Blackwell, 1984), 92-102, pp. 101-2.

thus have reason to reflect on just how grave the consequences will be if we don’t succeed in keeping the secret.

The consequences of being found out will be of two main types: those that are due to people’s or institutions’ disapproval of such acts and those which are independent of anyone’s disapproval. Consequences of the former type include the perpetrators’ being punished, those whose lives were saved feeling guilty, and the family of the dead person being outraged. All of these things are likely to happen if such an act is discovered in the actual world, and it’s probably utilitous that there is the sort of disapproval that would cause these reactions, since general disapproval of such acts helps prevent them from being performed on a wide basis, both by those who themselves disapprove of them and by those who fear the disapproval of others.

Knowledge that such acts were performed on a wide basis would be extremely disutilitous because of the way that it would affect people’s motivation to take care of their bodies. As I mentioned in Section V, each of us is naturally motivated to look out for our own interests (and generally for the interests of our family and friends), but, if it should happen that the efforts we take in view of procuring a pleasant future for ourselves have little chance of being effective, we will quickly lose a great deal of our motivation to act. The same thing will happen if the efforts we take in our own interest end up redounding solely to the benefit of others. And the effect will be even worse if, the more pains we take, the worse off we are likely to be.

This is the sort of system that would result, though, if healthy people routinely had their organs harvested to save the sick. The better care you took of yourself, the better candidate you would be for organ harvest, and, if you took little care of yourself, the negative effects of your neglect would be softened by the availability of others' organs. A decade or two of a healthcare system like this would very likely see a decrease in the average health of the population. In such a system, it would be to each person's benefit to use his body as hard as possible in procuring whatever pleasures he could because, in the case that some parts of his body gave out, someone else would pay most of the price. Under such a system, people would take less care of their bodies, there would be fewer organs worth transplanting, and total utility would likely be less than in a system where such transplants were disapproved of and not performed.

This is by no means to say that most of those who in the actual world need organ transplants are responsible for their plight, nor that most healthy people have done anything to deserve their health. The point is that, if we *removed* the incentives that currently exist, there *would* be many people who would get sick because of their lack of motivation to care for their own bodies. It thus seems like a bad thing for people to know that such transplants are widely performed in their society (if they are).

But will there be any negative consequences of people's finding out that *one* such transplant was performed, aside from punishment of the perpetrators and feelings of guilt and outrage? Will the discovery of one such transplant have any significant effect on people's expectations of non-interference with their bodies and their lives, significant enough to reduce their motivation to care for their health or invest in other long-term projects?

The question of whether one act can significantly increase others' expectations that similar acts will be performed in the future is of very general importance for the practice of utilitarianism. We've seen in Sections IV and V that it's clearly very utilitous for people in a society to have certain expectations about the acts that others will perform. It's less clear just how much reason this gives an individual to refrain from acting against these expectations on any particular occasion. In the introduction to this chapter, I quoted Williams expressing the belief that this sort of reason is much less strong than utilitarians have argued, certainly not strong enough to make it a decisive reason in these cases. Since this is a very common reaction to utilitarian arguments that the practice of utilitarianism largely follows our intuitions, it's important that utilitarians explain just how large a role this sort of reason plays in their arguments. This question is so important that I will dedicate Section VII to answering it.

VII. The probability of destroying useful expectations

It's tempting to think that performing one act couldn't possibly produce a significant increase in expectations that such acts will be performed in the future. It's especially tempting to believe that one act could not have this effect if the number of occasions on which people have already refrained from performing the act, and will likely refrain from performing it in the future, is very large. It seems clear that the number of occasions on which people refrain from performing transplants of the sort we're talking about *is* very large. And it also seems clear that there's a very large number of occasions on which people refrain from interfering with others' bodies and property, and on which they keep promises and tell the truth. Given this environment, it doesn't seem like one property-rights violation or one lie is going to change people's future expectations in any significant way. And perhaps neither will one unauthorized organ harvest.

I believe that our intuitions about the probable effects of such isolated acts are deceptive, however. To arrive at this intuition, we seem to do something like imagine some average person who hears about the incident, intuit that they would probably just brush it off as an aberration, not changing their behavior based on it, and conclude from the fact that the average person would not change their behavior that no one would. In reasoning this way, however, we fail to take into account a very important point. A single violation of one's expectations is likely to produce a very slight increase in uncertainty and a very slight decrease in motivation, even if these changes have no obvious effects on one's behavior in most cases. As the number of people affected and the number of occasions on which they will make decisions based on their expectations rise, the probability that some behavior will be affected by this slight change in expectations also rises. It becomes more and more likely that there will be some decisions made where the balance of reasons falls at just the right point for it to be tipped by this event.

The result, of course, will not be that the entire edifice of beneficial social expectations will crumble. The result will merely be that certain actions which would be utilitous won't be performed because agents won't feel confident enough about their expected utility. But the number of these utilitous actions that will not be performed will increase as the number of people whose expectations are slightly affected by the act increases, and also as the range of expectations affected by the act increases. Discovering that someone has been killed to have their organs harvested to save five others will likely affect not just people's expectations that their organs will be harvested (or that they'll get an organ transplant if they need one), but also their expectations that their other rights will be violated. With such a large number of our decisions dependent on our expectations that others will respect our rights to our bodies and property, the number of utilitous actions that won't be performed because agents won't feel confident enough about their utility could be quite high as a result of one unauthorized organ harvest. What needs to be taken into account in deciding whether to harvest the organs is not whether all expectations of respect for rights will disappear, but the collective disutility of all of these individual acts that will not be performed as a result of a *slight* change in expectations.

But, in addition to considering the effects of one isolated violation of expectations, we must also consider that our act may not be isolated, and that, in performing it, we may be contributing to a threshold effect. A threshold effect occurs when the effect of two or more acts' both being performed is greater than the sum of the effects of each act's occurring individually. It seems quite possible that people's psychologies are such that expectations about respect for their rights may change drastically when the violations of which they are aware reach a certain critical level. If this is the case, an agent contemplating killing one to save five needs also to calculate the probability that his performing this act will cause the threshold to be reached. This probability will depend on the probability of others' performing similar acts: If the probability is very high, then the threshold is likely to be reached even if this agent doesn't perform the act. If the probability is very low, the threshold will likely not be reached even if this agent performs the act.

But one should be very careful in making predictions about how many other similar acts are likely to come to light. Since anyone performing such an act will be striving for secrecy, one doesn't know how many are being performed, and so there is little basis for estimating how many are likely to be discovered! It's possible that one person's act's being discovered will lead to an investigation which will quickly uncover many more. And the fact that one person is seriously contemplating the act is itself good evidence that others are seriously contemplating it as well.

The fact that there are likely to be many other people thinking the way that we do doesn't mean, however, that a utilitarian should always refrain from acts that would have a strongly negative threshold effect. If, aside from the threshold effect, the acts have positive utility, then it will be optimal if the number of these acts just under the threshold is performed. If other people know this, then we can assume that they are all going to want just that number of us to perform these acts. Those of us who should perform them are those whose performance will have the highest utility. If it seems clear to everyone in which cases the utility of performing the act is particularly high, and that the number of such cases will be below the threshold, then it should also be clear to everyone that agents in those cases are justified in performing the act. The ones who should refrain are those agents whose situations are enough like enough other agents' situations that everyone can't be counted upon to know whether they fall to one side or the other of the threshold.[205]

But, returning to Transplant cases, it does not seem at all clear that the negative effects of others' finding out about one case of unauthorized organ harvest will be trivial. Given the number of people who will find out and the number of situations in which their choices could potentially be affected by a slight decrease in expectations of non-interference and increase in expectations of aid in case of organ

[205] Jan Narveson advocates this sort of response to threshold problems in "Utilitarianism, Group Actions, and Coordination or, Must the Utilitarian be a Buridan's Ass?" *Noûs* 10, no. 2, Symposium on Utilitarianism (May 1976): 173-94. For further discussion of threshold problems, see Lyons, *Forms and Limits of Utilitarianism*, Ch. 3; and Harry S. Silverstein, "Utilitarianism and Group Coordination," *Noûs* 13, no. 3 (Sept. 1979): 335-60.

failure, it seems likely that significant disutility will result from all the slight changes in expectations that will result. In addition to this, there is the possibility that other such acts may also come to light and that together these acts may cross a threshold with especially grave consequences.

VIII. Conclusion

With all of the considerations of Sections VI and VII laid out, I hope it will seem quite plausible that utilitarianism gives us reason not to harvest a healthy person's organs to save five other people. I hope it will be clear that the utilitarian case for refraining from these transplants rests not on just one of these considerations but on all of them put together. When we combine facts about our epistemic and motivational limitations with the particular facts of Transplant cases, they seem to provide strong reason to refrain from performing the transplants. And yet even if it's not entirely clear on which side the utility calculations come out, these considerations seem to cast enough doubt on the idea that performing the transplants will have positive utility that we have no reason to break from the generally utilitous practice of avoiding interference with others' bodies.

Similar sorts of considerations apply in other cases that have been used to try to show how utilitarianism has counterintuitive implications. For instance, we probably shouldn't frame and kill an innocent person just because we think doing so will save several other people from being killed by a rioting crowd. If we kill the one, he is sure to die, but we can't be sure that, if we don't kill him, several other people will be killed by the crowd, nor that killing him will reduce the crowd's violence. Maybe if we are repulsed enough by the idea of framing and killing an innocent person, we'll be motivated to think of a better way to stop the crowd from killing others. And in addition, if we frame and kill an innocent person, we reduce people's trust that they will not be killed in such a way, as well as their trust that they will be told the truth, and we encourage people to riot and threaten to kill others in order to get what they want.

It's also unclear that we should shoot one person just because someone we come upon in the jungle says that if we don't, he'll shoot twenty. If we shoot the gun, it's almost certain that the one person will die (and maybe the rest will still be shot). If we don't shoot, it's nowhere near certain that the twenty will all be killed. Maybe the man never intended to carry out his threat. Maybe the prisoners have a plan of escape. Maybe if we are repulsed enough by the idea of shooting the one, we'll be motivated to think of a better plan to save them all. And in addition, if we shoot the one and news of our action gets out, it could encourage others to make similar threats.

This second case is based on one described by Williams,[206] who in his version includes details such as that the prisoners are begging one to shoot, thus providing evidence that the man's threat ought to be believed. However, if we imagine being more and more certain that the twenty will die if we don't shoot the one, I think that

[206] Williams, 98-9.

our intuitions will start to support shooting the one. If we are being begged to shoot by all of those who best understand the situation—including by those who are in danger of being killed by us if we do as they say—then it is hard to see why we shouldn't do it. Perhaps we will still feel some reticence, due to our strong aversion to killing (an aversion that is generally utilitous) and some lingering doubts about the situation, but if we are entirely certain that our shooting one of these prisoners would avoid the deaths of the other nineteen and would have no further negative effects, I think many of us will feel that we ought to shoot the one, even if it's not easy.[207]

There are certainly other cases in which we feel so sure that the utility of violating rules of non-interference will be greater than the utility of not doing so that we feel certain we have reason to violate them. For instance, we all agree it's permissible to break someone's window if it's the only way to save ourselves from losing a limb. (We do think we ought to pay for the window, though, helping to ensure that self-interested bias doesn't keep us from finding a more utilitous solution to our

[207] Williams allows that shooting the one may be the right action, but he questions whether it is *obviously* the right action and whether in such cases there is not an important (though perhaps overridden) ethical consideration that utilitarianism does not account for: "a consideration involving the idea…that each of us is specially responsible for what *he* does, rather than for what other people do" (99). Utilitarianism does, however, justify a certain preoccupation with what *we* do rather than with what others do. The consequences of our actions which do not go through the intermediary of another agent are more certain than those that depend on the subsequent actions of someone else, and this justifies giving the former more weight in utility calculations.

Imagine, though, a case where the actions of another person are highly predictable, where there are so few inputs into her decision-making process that her reactions are just as sure as the outcome of pointing a loaded gun at someone's head and pulling the trigger. If someone's actions are this predictable, it's first of all difficult to see how she could rightly be called an "agent," and second of all, it seems clear that, if one is aware of this simple process by which her actions are determined, causing such a person to kill twenty people to avoid killing one person oneself is a grievous moral error.

problem.) Many people also agree that it's permissible to kill one person if we're almost certain to save a huge number of lives: thousands, or maybe even hundreds.

There's also an interesting class of cases in which we tend to agree that it's permissible to kill one person to save only five. Consider a case in which six people are adrift in shark-infested waters in a boat only big enough for five. The boat will sink and leave all of them to the sharks unless one person gets out. Most people aren't as horrified at the prospect of throwing one person overboard as they are at the prospect of harvesting a healthy person's organs.

This difference in our reactions may stem from the fact that all of the people in the boat seem to be in generally the same situation. If nothing is done, *none* of them will be able to survive, unlike in a Transplant case, where the healthy person could very easily go on living regardless of the fates of the other five people. The fact that all of the people in the boat case have ended up in equal peril seems to make us assume, in the absence of any further information, that none of them has any more of a right to live than the others. We assume that none of them has done anything more than the others to deserve to die, and that none of them has done anything more than the others to deserve to live.

But imagine now that this boat is the lifeboat of some larger craft which sank, and that one of the people in the lifeboat did everything he could to keep the larger craft afloat, while the other five ignored the impending danger. Surely the one who worked to keep the larger craft afloat should not be the one thrown overboard! Surely he has more of a right to live than the others, because of his efforts. The case suddenly resembles Transplant cases in the vehemence with which we reject the idea that the one ought to be sacrificed for the five. Our intuitions about who ought to be sacrificed for whom do seem to be affected by considerations of desert, exactly as the utilitarian justification for these intuitions says they ought to be, in order to harness the power of self-interest.

There’s no way to show, of course, that our intuitions always tell us to follow general rules in exactly those cases where following them is utilitous, and in no cases where it is not. I hope nonetheless to have provided some plausible reasons for thinking that our intuitions do on the whole advocate employing general rules to the same extent that the utility-maximizing decision procedure does. In any case, it does not seem that the deontological nature of our intuitions is nearly different enough from that of the utility-maximizing decision procedure to make our intuitions a reason to reject utilitarianism.

On the other hand, showing that our intuitions are consistent with the demands of specifically *hedonistic* utilitarianism will require a different set of arguments.

CHAPTER 7

THE PRACTICE OF HEDONISM

If utilitarianism as a whole is unpopular these days, hedonistic utilitarianism is even less so. Its few recent defenders include Leonard Katz,[208] T. L. S. Sprigge,[209] Torbjörn Tännsjö,[210] and Fred Feldman.[211] Uncoupled from utilitarianism, hedonism fares better, but not much. Usually when it's discussed, it's as no more than a quick set-up to praising a desire-based view of welfare. In James Griffin's 312-page book *Well-being: Its Meaning, Measurement, and Moral Importance*, the discussion of

[208] Leonard Katz, "Hedonism as Metaphysics of Mind and Value" (Ph.D. diss., Princeton University, 1986).

[209] T. L. S. Sprigge, *The Rational Foundations of Ethics* (London and New York: Routledge & Kegan Paul, 1988).

[210] Torbjörn Tännsjö, *Hedonistic Utilitarianism* (Edinburgh: Edinburgh University Press, 1998).

[211] Fred Feldman, *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy* (Cambridge: Cambridge University Press, 1997); and *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism* (Oxford: Clarendon Press, 2004).

hedonism occupies less than three pages.[212] Granted, there are a few notable exceptions to the anti-hedonistic trend—in addition to the hedonistic utilitarians mentioned above, Rem B. Edwards[213] and Roger Crisp[214]—but their presence has not prevented general opinion from reaching the point at which refuting hedonism can seem too trivial an accomplishment even to justify an entire article: Elijah Millgram, after presenting some arguments against hedonistic utilitarianism in "What's the use of utility?", worries that his readers will wonder whether his target "was not in fact a straw man," and proceeds to address his attention to preference utilitarianism, because what he wants us "to take away from this exercise is a more important lesson than: that an already-discredited view has been further discredited."[215]

The facility with which so many contemporary philosophers dismiss hedonism is rather surprising, given the serious attention it's historically received. Three Platonic dialogues—the *Protagoras*, the *Philebus*, and the *Republic*—seriously discuss it. So does Aristotle in *Nicomachean Ethics*.[216] The Epicureans prominently defended the

[212] James Griffin, *Well-being: Its Meaning, Measurement, and Moral Importance* (Oxford: Clarendon Press, 1986), 7-10. For another quick dismissal of hedonism, see Will Kymlicka, *Contemporary Political Philosophy* (Oxford: Clarendon Press, 1990), 12-14.

[213] Rem B. Edwards, *Pleasures and Pains: A Theory of Qualitative Hedonism* (Ithaca: Cornell University Press, 1979).

[214] Roger Crisp, "Hedonism Reconsidered," *Philosophy and Phenomenological Research* 73, no. 3 (November 2006): 619-42; and *Reasons and the Good* (Oxford: Oxford University Press, 2006), Ch. 4.

[215] Elijah Millgram, "What's the use of utility?" *Philosophy and Public Affairs* 29, no. 2 (Spring 2000): 113-36, pp. 126-27.

[216] Aristotle, *Nicomachean Ethics*, 7.11-14, 10.1-5.

view, even as the Stoics went to pains to argue against it.[217] And much more recently, hedonism was the preferred view of the British empiricists, defended by Hobbes, Locke, Hume, Bentham, and Mill.[218,219]

The causes of the twentieth-century decline of interest in hedonism are not entirely clear. Crisp suggests that the first blow was dealt by Mill's attempt to avoid counterintuitive consequences for hedonism by appealing to a distinction between higher and lower pleasures, a distinction which appeared to be "either an abandonment of hedonism or incoherent."[220] Crisp also mentions the subsequent influence of Moore's criticisms of hedonism in *Principia Ethica*.[221] I believe twentieth-century enthusiasm for behaviorism and propositional-attitude psychology played a large part in marginalizing hedonism. Daniel Kahneman and Carol Varey second this view, regretting that "[t]he definition of utility in terms of choices still rules the sciences of decision, although operationalism and behaviorism have largely lost their hold on

[217] On the Epicureans, see Cicero, *De Finibus Bonorum et Malorum*, 1.30-54. On the Stoics, see Diogenes Laertius, *Lives of Eminent Philosophers*, ed. and trans. R. Hicks, rev. H. Long (Cambridge, Mass.: Harvard University Press, 1972), 7.85-6.

[218] Thomas Hobbes, *Human Nature: or the Fundamental Elements of Policy*, 7.3; John Locke, *An Essay Concerning Human Understanding*, 2.20.3; David Hume, *Enquiry concerning the Principles of Morals*, app. 2.10; Hume, *A Treatise of Human Nature*, Book 2, Part 3, Section 9; Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, 1.1; John Stuart Mill, *Utilitarianism*, reprinted in Albert William Levi, ed., *The Six Great Humanistic Essays of John Stuart Mill* (New York: Washington Square Press, 1963), 241-308, especially Ch. 2.

[219] I owe these historical references to Crisp, 619.

[220] Crisp, 619-20.

[221] G. E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1903), Ch. 3.

psychology."[222] Currently most important to hedonism's unpopularity, however, seems to be the "experience machine" thought experiment, published by Robert Nozick in 1974.[223] References to Nozick's experience machine tend to dominate what contemporary discussions of hedonism exist, making it seem likely that if the continuing unpopularity of hedonism can be traced to any single cause, it's the fact that many people feel repulsed by what they suppose is a hedonistic paradise: a world of people enjoying the most pleasurable experience possible, provided by a machine that directly stimulates their neurons.

Since the goal of this chapter is to make hedonism more plausible to contemporary readers, I'm going to spend a large portion of this chapter discussing the implications of Nozick's thought experiment. In preparation for this discussion, I'll spend Section I making some preliminary points about the practice of hedonism, showing how a hedonist not only has to take into account the intrinsic goodness and badness of pleasure and pain[224] but also their instrumental value as indicators of future prospects for pleasure and pain. I'll show how it is that, even for a hedonist, there are fitting and unfitting things in which to take pleasure.

In Section II, I will turn to the experience machine. I will argue that our negative feelings about being hooked up to the experience machine are *not* in fact

[222] Daniel Kahneman and Carol Varey, "Notes on the psychology of utility," in Jon Elster and John E. Roemer, eds., *Interpersonal Comparisons of Well-being* (Cambridge: Cambridge University Press, 1991), 127-86, p. 128.
[223] Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), 42-45.
[224] In this chapter, I will use the term 'pain' to refer to all negative experiences, in keeping with the traditional hedonistic usage.

evidence against hedonism, and I will argue for this conclusion by two different routes. First, I will argue that there are actually hedonistic reasons for rejecting life hooked up to the experience machine, due to the instrumental benefits of being in contact with the world outside of our experiences, as described in Section I. (I will also show that, if we look at cases where these hedonistic reasons don't apply, and flesh out why they don't apply, our intuitions about hooking up to the experience machine in these cases become much less negative.) Second, I will argue that, even if we do value things besides pleasant experience as ends in themselves—and so would perhaps *never* consent to spend life hooked up to the experience machine, even in worlds where hedonistic reasons for not doing so didn't apply—the fact that we value these other things as ends in themselves is still not evidence against hedonism, for two reasons: In Section III, I will discuss the fact that the experience machine argument depends on the premise that there is a connection between what we desire or feel to be valuable and what is objectively valuable, and I will argue that this premise needs to be justified before this can be a conclusive argument. In Section IV, I will argue that, in the actual world, valuing certain things besides pleasure as ends in themselves actually serves hedonistic ends. That is, hedonism actually justifies our seeing these other things as intrinsically valuable, and so our "non-hedonistic" values actually accord with the practical demands of hedonism.

Once I've finished arguing that our feelings about the experience machine are not at odds with hedonism, however, I will close the chapter by addressing an intuition that many people may have that is indeed contrary to the form of hedonism that I've developed in this dissertation. This is the intuition that pleasure had by humans is of greater intrinsic value than that had by non-human animals. I will explain why we ought to mistrust this intuition and seek to reform it, rather than taking it as a reason to reject hedonism.

I. Pleasure and pain as indicators

When many people think of what it would be like to live as a hedonist (or at least as an egoistic hedonist), they conjure up images of people who deny themselves no opportunity for pleasure, living in perpetual indulgence of their appetites for food, drink, drugs, and sex. The fact that most people who actually attempt such a lifestyle either abandon it rather quickly or die prematurely should teach us something about the actual value of such a life. Constantly seeking to procure the most immediate pleasure and to banish all pain is not in fact a good method for producing the most pleasure and the least pain in the long term. One important reason for this is that our bodies simply cannot support constant stimulation and must be given time to rest and replenish themselves. Constant stimulation—especially in the form of drugs—interferes with the body's sustaining of its most basic functions and leads to its ultimate breakdown. Furthermore, continuous intense pleasure prevents us from consciously attending to bodily needs because it disconnects the state of our mind from the state of our bodies. Drugs can make us feel that everything is going wonderfully when in fact our bodies are suffering acutely. By directly inducing intense pleasure, we drown out the signals our bodies otherwise send as a reminder that they require attention: that they require nutrients, water, and sleep, for example. The problem is demonstrated by rats that have been taught that they can electrically stimulate the "pleasure centers" of their brains by pressing a lever. These rats choose pressing the lever over any other activity, including eating, drinking, and copulating, and if allowed, they will press the lever until they collapse.[225]

Those who assume that a hedonistic lifestyle would be one of constant pleasure without pain have missed the very important point that, even for a hedonist, pleasure and pain are not only ends in themselves, but also serve as means. Pleasure and pain have an *instrumental* role to play, as indicators of the instrumental goodness or badness of other things. It's interesting that, while the role pleasure and pain play in indicating the goodness or badness of other things has been noted by many philosophers, their having this indicative role is often taken as evidence *against* hedonism rather than for it.[226] Millgram writes, "Hedonists assume that because desires and goals change in response to experienced pleasure and displeasure, these must be the actual goals. But this view is naive: pleasure and displeasure are indications and signs of desirability we use in determining what our goals should

[225] Edwards, 60.

[226] See, for example, Millgram, "What's the use of utility?"; Millgram, *Practical Induction* (Cambridge, Mass.: Harvard University Press, 1997), Ch. 6; and Troy Jollimore, "Meaningless Happiness and Meaningful Suffering," *Southern Journal of Philosophy* 42, no. 3 (Fall 2004): 333-47.

be."[227] Millgram and others ignore the fact that even hedonists can accept that pleasure and pain indicate the desirability of other things, a desirability which consists in those things' tendencies to produce future pleasure or reduce future pain.

In fact, pleasure and pain are instrumentally valuable not just because they can *indicate* prospects for their own promotion or avoidance but because they can actually motivate us to take *action* towards their promotion or avoidance. We could find other ways of informing ourselves about the state of our bodies—by mechanically monitoring our vital signs, our blood sugar level, etc.—but simply having information about what we ought to do to maintain our health is not enough to get us to do it, especially when we are enjoying intense pleasure or have it as an immediate prospect.[228] The insufficient motivational power of mere information about the harm our actions are doing our bodies has been documented by Paul Brand, who spent years developing artificial sensing systems for victims of leprosy.[229] Since leprosy makes a patient insensitive to pain in their hands and feet, Brand developed gloves and socks containing pressure sensors capable of indicating when some action the patient was taking was having a harmful effect on their body. Unfortunately, these indications that they were performing an action such that it could lead to the destruction of a finger or toe did not stop patients from performing the action. They were not sufficiently

[227] Millgram, *Practical Induction*, 117.

[228] Millgram makes a similar point in "What's the use of utility?", 125-6.

[229] Paul Brand and Philip Yancey, *Pain: The Gift Nobody Wants* (New York: Harper Collins, 1993), especially pp. 194-96.

motivated by the evidence about their future well-being, and the project ultimately failed.

What is needed in such cases is not just the knowledge that one's present behavior will have future negative consequences but a *present* negative consequence—such as a negative normative quale—that directly motivates one to cease the harmful behavior.[230] Brand reports the comment of a colleague, Professor Tims, who said to him, "Paul, it's no use. We'll never be able to protect these limbs unless the signal really hurts. Surely there must be some way to hurt your patients enough to make them pay attention."[231] In fact, Brand and his colleagues did try an alternative system which responded to danger to the limbs by inducing pain in a still sensitive part of the body. The problem with this system was that patients preferred to turn it off rather than heed its warnings! The moral is that feeling some present pain is crucial to motivating us to avoid future negative consequences, and that, in order for pain to be so motivating, there has to be no easier way to avoid the pain than to avoid the harmful behavior.

Negative normative qualia are not only useful in motivating us to attend to the state of our bodies but also in motivating us to take care of our social lives. It is crucial to our long-term happiness that we have strong ties to others in our society and

[230] The question of whether normative qualia are the only mental states that can have this motivational effect, or whether they have this causal property essentially, is not one I'm going to get into here. What's important is that as a matter of fact they have a connection to motivation that most other mental states lack.

[231] Brand and Yancey, 194.

particularly strong ties to a few individuals. We rely on others throughout our lives for food, housing, medical care, and protection, as well as for the pleasures that come directly from close personal relationships. Thus it's essential to our happiness that we have the capacity to live in community with others, and this requires sensitivity to the needs of others and to the general requirements of membership in a group.

For instance, it's important that when we hurt a friend's feelings, we feel emotional pain. Without it, we might very well think little about what we've done until the time comes when we need something from our friend. By that point, we may have permanently lost the friendship. Feeling pain over a rift in the relationship motivates us to make amends in a timely manner, by providing a foretaste of all of the future pain a loss of the friendship might entail. We are also benefited by our ability to experience shame, embarrassment, and discomfort at the disapproval of others. These feelings motivate us to please others, and their pleasure in turn causes them to have positive personal associations with us and to see us as someone who should be helped in the future, because we've demonstrated our potential for a mutually beneficial relationship.

Now I've so far been talking about the way in which negative normative qualia can signal to us things we need to change in order to spare *ourselves* future problems. But of course if we are hedonistic *utilitarians*, and not just hedonistic egoists, for example, than we will believe that our actions ought to be such as to maximize the balance of pleasure over pain not just in our own lives but across *all* people's lives. If hedonistic utilitarianism is true, it will be wrong to abandon oneself to one's own sensual or drug-induced delights if this causes one to neglect the interests of those around one, where these latter are important enough to outweigh one's personal pleasures. And in fact, others' interests have a high probability of outweighing the pleasure one would gain from ignoring them. This is partly because some of the suffering in the world is just so acute (and the cost of alleviating it so low in comparison). But even in cases where there is no great suffering that one could relieve, there is the fact that one could take pleasure in something that is not self-contained like a drug-induced high, but rather is an act which simultaneously brings pleasure to someone else. The choice is not always between *x* units of pleasure for me and *y* units of pleasure for someone else. Instead, it's often between (1) *x* units of pleasure for me and (2) *y* units of pleasure for someone else *plus* the amount of pleasure I'll take in bringing that pleasure to them.

One of the most effective general strategies a utilitarian has for achieving the greatest overall balance of pleasure over pain is encouraging states of affairs in which different people's pleasure is interconnected.[232] The more people there are who take pleasure in others' pleasure, the better the world will be. Those people will have an

[232] Mill makes this point in Chapter 2 of *Utilitarianism*, writing that "utility would enjoin, first, that laws and social arrangements should place the happiness or (as, speaking practically, it may be called) the interest of every individual as nearly as possible in harmony with the interest of the whole; and secondly, that education and opinion, which have so vast a power over human character, should so use that power as to establish in the mind of every individual an indissoluble association between his own happiness and the good of the whole—especially between his own happiness, and the practice of such modes of conduct, negative and positive, as regard for the universal happiness prescribes…."

incentive to bring pleasure to others, and every time they do, that pleasure will be multiplied by their own pleasure. Because a single individual could not meet all of his or her own needs in any case, it is crucial that each of us look out for the needs of others that they cannot best meet on their own, and the way that we motivate this sort of behavior is through making our pleasure and pain dependent on theirs, at least to some degree.

Thus we have seen that hedonism can respond to a common objection leveled against it: that it does not take into account the fact that there are both fitting and unfitting things in which to take pleasure. Hedonism can agree that pleasure ought to be taken in things such as relationships, beauty, and a job well done, and that it ought not to be taken in hurting others. It can also agree that pleasure should not regularly be derived from drugs or from electrically stimulating the pleasure centers of one's brain. And rather than taking these as brute facts, hedonism can explain *why* certain things are the ones it's fitting to take pleasure in: because taking pleasure in these things leads to more pleasure overall, while taking pleasure in other, unfitting things can end up causing a lot of pain, or preventing one from producing greater pleasures.

II. Why the experience machine is a bad idea

Now that I've explained the basic reasons that a hedonist has to want her pleasure and pain to reflect future prospects for pleasure and pain (whether just for herself or also for others), we can turn to applying these considerations to the case of

Nozick's experience machine. In an attempt to show that experiences are not the only valuable things in life, Nozick proposes the following thought experiment:

> Suppose there were an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences? If you are worried about missing out on desirable experiences, we can suppose that business enterprises have researched thoroughly the lives of many others. You can pick and choose from their large library or smorgasbord of such experiences for, say, the next two years. After two years have passed, you will have ten minutes or ten hours out of the tank, to select the experiences for your *next* two years. Of course, while in the tank, you won't know that you're there; you'll think it's all actually happening. … Would you plug in? ... We learn that something matters to us in addition to experience by imagining an experience machine and then realizing that we would not use it.[233]

Our distaste for living hooked up to an experience machine has been cited by Nozick, and by many philosophers after him,[234] as evidence that hedonism is false, that things besides pleasant experience are important to making one's life as good as possible.

There are at least two basic problems with taking our feelings about the experience machine as evidence that hedonism is false. First of all, we can't simply

[233] Nozick, *Anarchy, State, and Utopia*, 42-44.

[234] See, for example, Griffin, 9-10; David Brink, *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press, 1989), 223-24; Stephen Darwall, "Self-Interest and Self-Concern," *Social Philosophy and Policy* 14 (1997): 158-78, pp. 162, 178; L. W. Sumner, *Welfare, Happiness, and Ethics* (Oxford: Clarendon Press, 1996), 94-98; John Finnis, *Natural Law and Natural Rights* (Oxford: Clarendon Press, 1980), 33; Finnis, *Fundamentals of Ethics* (Washington, D.C.; Georgetown University Press, 1983), 37-42; Garrett Thomson, *Needs* (London: Routledge & Kegan Paul, 1987), 41; and Robin Attfield, *A Theory of Value and Obligation* (London: Croom Helm, 1987), 33.

assume that *our desiring* or *our valuing* things besides pleasant experience means these things are *objectively* valuable, that our having them makes us objectively better off. The connection between our desires and objective value needs to be argued for, as I'll discuss in Section III. But even if we think we are justified in trusting our intuitions about life hooked up to the experience machine to tell us whether it's objectively valuable, that doesn't seal the deal against hedonism, since hedonism may *also* tell us to reject life hooked up to the experience machine, for reasons along the lines of those described in Section I.

Consider first egoistic reasons a hedonist has not to live his life hooked up to the experience machine. Being hooked up to the experience machine isolates him from any information about his future welfare and renders him incapable of any action that could improve his future balance of pleasure over pain. He is completely at the mercy of the machine, those who are operating it, and anyone else who might come into contact with him or the machine. He must trust those outside of the machine to look out for his interests as well as he could himself if he were living disconnected from the machine. And he must trust that, if something goes wrong with the machine or with his brain, such that he is no longer having pleasurable experiences but painful ones, he will be unhooked from the machine. The unlikelihood that any people or machines could be as good at looking after our interests *without* our conscious participation as *with* it gives our negative feelings about living hooked up to the experience machine one very strong hedonistic justification. And while we may not consciously be reflecting on all of this when we think about the experience machine, it seems likely that these kinds of concerns are operating somewhere in the background, giving us a general uneasiness about it.

Consider, in addition, that living hooked up to the experience machine wastes all of the potential we have to improve the lives of others. Of course, we can imagine that there might be some individuals who have dispositions such that their interacting with the external world causes more pain than pleasure. Such people have hedonistic reason *not* to be in contact with the external world, and isolating them in an artificial world of pleasant experiences may be just the right thing to do. (It would certainly be better for them than putting them in prison, although it might not be a very effective deterrent for bad behavior!) It seems like intuitions about the experience machine are probably going to be less negative in a case like this, too. But there are nevertheless many others of us who have dispositions such that our interaction with the external world normally benefits others, and this could help to justify (and cause) our negative feelings about living hooked up to the experience machine.

Now we should note that Nozick's intention in giving the experience machine example is to focus our attention on what makes a life good *for the person living it*. That is, he wants to rule out justifications for our negative feelings about the experience machine in terms of other-regarding reasons. I'm not sure that we can easily isolate our intuitions about what is a good life *for the person living it* from our feelings about what makes lives good all things considered. (This goes back to the difficulty of distinguishing our intuitions about intrinsic and instrumental goodness, which I discussed in Chapter 4, Section I.) In an effort to do this, however, Nozick tells us to consider not just a world where *we* are hooked up to an experience machine, but one where *everyone* is.[235] The assumption is that, since everyone in such a situation would be capable of automatically obtaining as much pleasure as possible, there would be no need for any individual to look after the needs of others, or after his own future interests. Everyone would be equally taken care of by the experience machine. Thus a hedonist whose intuitions were previously affected by concern for others should see no problem with hooking up to the experience machine in a world like this.

But let's examine such a situation in detail. If there is to be absolutely no need for anyone to have knowledge of what is going on in the non-virtual world, a huge number of things are going to have to be taken care of. Everyone's body will have to be fed and hydrated, and its health otherwise maintained. This will require that the production of food, water, and energy, as well as the provision of medical care, all be entirely automated. It will have to be the case as well that the experience machine (not to mention all of the machines providing food, water, energy, and medical care) will never break or run out of fuel or crash due to software glitches. They must never run out of stimulating experiences, and, if we are utilitarians, then we will believe that they must also see to human reproduction—or to keeping the same bodies alive

[235] Nozick, *Anarchy, State, and Utopia*, 43.

forever—in order to ensure that as much pleasure as possible continues to be produced as long into the future as possible.

What all of this means is that, in order for it to be clearly a good idea for us all to plug into the experience machine, we will have had to create machines which are better problem-solvers than we are. We will have had to create machines which are better at dealing with all the unpredictable aspects of life—weather, climate change, and fuel shortages, to name a few—than we ourselves. (Not to mention the fact that, if there are some people who choose not to live hooked up to experience machines, our machines will also have to manage our relationship to those people, preventing them from harming us and our cadre of machines by engaging in diplomacy, and perhaps even fighting wars.) It is hard to believe that we could create machines so superior to ourselves, not just in making numerous computations, but in dealing creatively with new situations. It seems that there is very little chance that we would ever be better off putting ourselves permanently at the mercy of a machine we've created than in dealing with the world face-to-face.

We should acknowledge, though, that hooking up to the machine doesn't have to be an irrevocable decision.[236] We could be disconnected from it from time to time, in order to make an evaluation of our situation in the external world and deal with any problems. One difficulty with this solution, however, is that problems might arise that

[236] Nozick's original example mentions that one could come out of the machine from time to time, and Tännsjö has reiterated this point (111-12).

need to be solved before our next programmed “wake-up” date. Even if we program the experience machine to wake us up whenever there’s a problem it can’t solve, we still have to rely on the machine’s ability to recognize such problems in time for something to be done about them. And even if we assume that we can manage to get ourselves woken up whenever there’s a problem at hand that the machine can’t solve, it’s still unclear that *we* will be able to solve the problem if we haven’t been living in contact with the external world. We won’t have had all of the day-to-day experiences that normally teach us about our environment and make us good problem-solvers within it.

In sum, hooking up to the experience machine for long stretches of time would be a bad idea in any world with problems as complex as those of the actual world. We have hedonistic reasons to put our problem-solving capacities to work by regularly interacting with the external world, in order to protect both our own future interests and those of others. It thus appears that our intuition about the badness of spending one’s life hooked up to the experience machine is not at odds with the demands of hedonism in the actual world or in any worlds similar to it in complexity.

This doesn’t prove, of course, that our negative intuition about the experience machine is caused by these hedonistic reasons. For all we’ve seen so far, it could be that our intuition is actually responding to the intrinsic goodness of having true beliefs or of having real interaction with other human beings, and that these reasons just happen to give us the same negative verdict about the experience machine that hedonism gives us, in worlds close to the actual one. It’s hard to know whether this is the case, given that the reasons for our intuitions aren’t nearly as obvious to us as the intuitions themselves. It does seem that even the mere agreement of our intuitions with the demands of hedonism in worlds close to this one ought to do *something* to increase the palatability of hedonism, especially if we doubt the reliability of our intuitions in cases very different from those that have shaped their development. But I think we may be able to show something more.

Let’s try again to imagine a world in which there are no hedonistic reasons—either other-regarding or self-regarding—against hooking oneself up to the experience machine. This is going to have to be a world in which human beings have done absolutely all we ever can to solve problems of hunger, health, politics, war, resources, technology, and natural disasters. For this to be true—and for us to know that it’s true—we will have to have a close to exhaustive knowledge of our environment and of human biology and psychology. And yet, if this is the case, what is it one plans to do if one refuses to plug into the experience machine? We know how to satisfy every one of our needs as soon as we have it. There are no experiences to have that haven’t already been had and catalogued in the experience machine. There’s no one to dedicate one’s life to helping. What things remain to be learned about the universe are going to be things we know, *a priori*, can have no effect on improving the quality of anyone’s life—things like the exact number of atoms there are, or historical facts that are so specific that they don't reveal to us any general truths about the world or human nature that we weren't already aware of.

Faced with a choice between living in a world where all that remains to be done or discovered are things of this little consequence, and plugging into the experience machine where at least one has the *experience* of doing and discovering worthwhile things, the experience machine doesn't look so bad. Once we actually imagine what the world would have to be like in order for contact with the external world not to have instrumental value, hooking up to the experience machine in that kind of case seems quite reasonable! Our intuitions about the value of a life hooked up to the experience machine thus seem at least roughly to parallel the demands of hedonism even in cases like this, which differ substantially from the actual one.

We should note, too, that even in the actual world, there are cases in which there are no strong hedonistic reasons to refrain from cutting oneself off from the external world *for a limited period of time*, and many of us do just that. Many of us participate on a daily basis in something very much like the experience machine: by watching television, watching movies, playing video or computer games, reading books, and sleeping. And we disconnect from the "real world" for even longer periods of time when we spend a day at the spa or a week on a tropical beach. These activities avoid some of the major disadvantages of plugging into the experience machine, because (1) even while we are engaged in them, they allow us to remain at least somewhat alert to things in the real world that may require our attention, and (2) they absorb only a portion of our lives and thus still allow us to remain agents capable of promoting our own happiness and that of others. On the other hand, to the extent that these activities do not meet these criteria, we usually do consider engaging in them worrisome. Here, too, our intuitions about maintaining connection to the external world seem to be generally consistent with hedonistic reasons.

III. Our valuing something vs. its having objective value

It may be, however, that our intuitions do not perfectly track these reasons. Perhaps some people will feel that, even in a world in which all problems are solved and we know everything there is to know about everything and everyone, there would still be some value to living in contact with the external world rather than plugging into the experience machine. Perhaps some people will insist that real relationships with others are intrinsically valuable even in such a world, and that the value of having these relationships gives us *some* reason not to plug into the experience machine, even if it's outweighed by the need to escape from boredom.

I think all of us probably have some intuitions which are at odds with hedonistic reasons, at least in worlds very different from the actual one, and this is because I think many of us do place intrinsic value on things besides pleasurable experience. For instance, we value not just having the *experience* of having a family but *actually having* a family. We think it would be a terrible thing if instead of interacting with flesh-and-blood family members as we think we're doing, we were only interacting with computer simulations. And we don’t just want to *feel* as though we’re engaged in a fulfilling romantic relationship. We want there actually to be another person whom we love and who loves us. While we may not be able to tell the difference between the independent existence of these things and some future computer simulation of them, that doesn’t mean we can’t think their independent existence is important. We care not just about how we think the world is, but about how the world is in itself.

Of course, a hedonistic utilitarian should also care about how the world is outside of his experiences. A hedonistic utilitarian should care about the experiences of all other experiencing beings. It is most definitely important to the utilitarian that his romantic partner be real and not a computer simulation. Something of real value is missing if one is engaged in a merely virtual relationship: the positive experiences that would otherwise be given to one’s partner.

However, we can stipulate that there are as many experiencing beings alive as the world’s resources will support and that they are all guaranteed to be forever maintained in an optimal state of pleasure by the experience machine. If this is the case, then one’s desire to be in an actual interactive relationship with another human being cannot have a hedonistic justification. The world already contains the maximum possible amount of pleasure, and thus carrying on an actual interactive relationship with another person can’t increase this amount. Yet I imagine some people feel the addition of interactive relationships would make such a world better. And they conclude from this that hedonism is wrong to say that only experiential states are valuable.

To respond to this objection, we need to draw a clear distinction between the act or attitude of valuing and the having of objective value. *Valuing* is what people do; it's an activity or disposition which involves desiring something and approving of that desiring. It could be characterized, à la Frankfurtienne, as desiring "whole-heartedly." *Having objective value*, on the other hand, is an objectively normative property of an object, event, or state of affairs, such as a positive normative quale.

The distinction between valuing and having objective value is frequently ignored by philosophers who invoke our desires for things besides experiences to refute hedonism.[237] Nozick himself ignores the distinction in his original statement of the experience machine argument. Matthew Silverstein points out the way in which Nozick's discussion plays on an ambiguity in the meaning of the word "matters."[238] Something can be said to "matter" in the sense of being objectively valuable, or it can be said to "matter *to*" an individual, in the sense of their desiring it or valuing it. (To make things even more complicated, something can also "matter to" an individual in an objective way: e.g., it might matter to a particular individual's welfare that they quit using drugs, regardless of whether they actually desire to quit.) These senses are not

[237] See, for example, Griffin, 7; and Kymlicka, 14. This criticism of Griffin, Kymlicka, and Nozick is made by Jason Kawall, "The Experience Machine and Mental State Theories of Well-being," *Journal of Value Inquiry* 33 (1999): 381-7, pp. 384-5.

[238] Matthew Silverstein, "In Defense of Happiness: A Response to the Experience Machine," *Social Theory and Practice: An International and Interdisciplinary Journal of Social Philosophy* 26, no. 2 (Summer 2000): 279-300, p. 286.

clearly distinguished in Nozick's discussion, leaving implicit his assumption that what someone values is evidence for that thing's being objectively valuable.[239]

In a later discussion of the experience machine, Nozick clarifies that he does indeed mean to be taking our desires as indicating objective value:

> Notice that I am not saying simply that since we desire connection to actuality the experience machine is defective because it does not give us whatever we desire—though the example is useful to show we *do* desire some things in addition to experiences—for that would make "getting whatever you desire" the primary standard. Rather, I am saying that the connection to actuality is important whether or not we desire it—that is *why* we desire it—and the experience machine is inadequate because it doesn't give us *that*.[240]

But even though Nozick clarifies his position in this way, he doesn't offer an argument for his claim that our desire for a connection to actuality reflects its objective importance.

Nozick probably means to rely on something's *seeming* valuable (and thus provoking desire) as *prima facie* evidence for its actually *being* valuable. This would be fine if something's being valuable was always reducible to facts about our perception of it (as is the case with the intrinsic goodness and badness of phenomenal qualities). But if the value of being connected to the real world is supposed to be independent of any perception of ours, then our perception's being *evidence* for its value must be established by evidence of a correlation between our perceptions and

[239] Of course, this assumption is not restricted to opponents of hedonism. Mill famously employed the premise that "the sole evidence it is possible to produce that any thing is desirable is that people do actually desire it" in an argument *for* hedonism (*Utilitarianism*, Chapter 4).

[240] Nozick, *The Examined Life: Philosophical Meditations* (New York: Simon and Schuster, 1989), 106-7.

perception-independent value. That is, we have to have some perception-independent method of determining which things are valuable to confirm the reliability of our perceptions of value. I'm very pessimistic about the possibility of finding such a method, and that's why I endorse the view that the only thing intrinsically normative is phenomenology itself, the intrinsic qualities of which are identical to the qualities we experience them as having.

IV. Why taking something to be intrinsically valuable could be instrumentally valuable

But can hedonism give any explanation or justification for why, if our perceiving certain things to be intrinsically valuable is not reflective of their objective intrinsic value, we nevertheless perceive them as such? In fact it can. I gave a psychological explanation for such perceptions in Section V of Chapter 3, where I explained that if the perception or thought of an object always produces a positive normative quale, this can cause us to see the goodness of this quale as being in the object rather than in the phenomenology, such that we don't make a distinction between the object and the phenomenology, and we pursue the object as if it were an end in itself. (We might also offer psychological and evolutionary explanations as to why normative qualia come to be associated with the specific objects they do, though I won't do that here.) In addition to hedonistic *explanations* for this phenomenon, however, there is a hedonistic *justification* for seeing certain things besides pleasure as intrinsically valuable.

Consider first that many of the things that we take to be intrinsically valuable besides pleasure have instrumental value, but an instrumental value that is optimally promoted if we think of them as intrinsically valuable. I believe this is true of things such as relationships, knowledge about the material world, and beautiful objects or works of art. None of these things is objectively intrinsically valuable, but they are all so instrumentally valuable that thinking of them as intrinsically valuable rarely leads us astray. Thinking of them as intrinsically valuable can actually be a way of saving ourselves the time of thinking about all of the various ways in which they are instrumentally valuable, on every occasion when we have to make a choice involving them. Thinking of certain instrumental goods as intrinsic goods can be an efficiency whose value over time outweighs whatever negative consequences result on the isolated occasion on which the object happens not to be instrumentally beneficial.

But this is not the only reason to think of relationships, knowledge, and art as intrinsically valuable. Because of certain facts about the way relationships, knowledge, and art are produced, thinking of them as valuable in themselves can be more conducive to their production than thinking of them as mere means. This is one aspect of what is often referred to as the "paradox of hedonism": the fact that we often maximize pleasure only if we take something else as our goal.

Take the development of a relationship. If one is constantly preoccupied with calculating just how much future pleasure a young relationship promises, one will not be capable of the sort of self-abandonment that the creation of a strong emotional tie requires. The greatest pleasures of a close personal relationship depend on one's abandoning the project of comparing the relationship to others on instrumental grounds and embracing the present relationship as in itself worthy of nurturing. This is not to say that we should never take into account whether a relationship is actually making us happy. It's just to say that such concerns cannot be our focus if the relationship is to develop to its full potential.[241]

The situation is similar with respect to the pursuit of knowledge. While knowledge is ultimately valuable because of the way it contributes to increasing the balance of pleasure over pain, history seems to show that the single-minded pursuit of truth about the world, with no concern for what technological value it may have, may ultimately be very fruitful technologically. This may be simply because, when we don't know certain things about the world, we also don't know whether knowing those things would be useful or not. It also seems to be the case that a love for the search for truth can sustain long years of pains-taking research that someone more directly interested in technological (and monetary) pay-offs would never have endured. And yet those long, seemingly fruitless years can ultimately produce a discovery of immense technological import. Sometimes, in order to achieve a far-off goal, it's necessary that we be presently motivated by an attitude that what we are doing here and now is itself intrinsically valuable.

[241] On this topic, see also Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," *Philosophy and Public Affairs* 13, no. 2 (Spring 1984): 134-71, p. 142.

Artists, too, it seems, are often hindered if they consider what function—aesthetic or otherwise—their work will ultimately serve, or what fame it will bring them. We get more creative and interesting pieces if the artist pursues art for its own sake and the results are judged afterward.[242]

For these reasons, a hedonist ought to encourage thinking of relationships, knowledge, and works of art as intrinsically valuable. Whenever the instrumental value of a thing is great enough and consistent enough, it saves time to treat it as an intrinsic good. And, perhaps more importantly, being thought of as an intrinsic good may in fact be the only way that certain things can develop to the point of being instrumentally valuable. This doesn't mean that thinking of these things as intrinsically good will *always, in every situation* produce the greatest balance of pleasure over pain. There may be situations, especially in worlds far from the actual one, in which these things are much less instrumentally valuable (or other things are much more valuable), and taking them to be intrinsically valuable will cause us to make a less-than-optimal decision. But as long as their instrumental value is sufficiently high in the actual world, these situations will be rare here. And if one can't stop treating these things as intrinsically valuable in a few rare situations without greatly diminishing one's ability to promote their instrumental value in all the rest

[242] Sidgwick also discusses the hedonistic benefits of focusing on knowledge or art for its own sake, although he tends to focus on increases in the pleasure immediately enjoyed in learning or creating art and not on its further instrumental effects. See *The Methods of Ethics*, 7th ed. (London: Macmillan, 1913), 49. For further discussion of the paradox of hedonism by Sidgwick, see the rest of Book I, Ch. IV, as well as pp. 403, 405-6, in Book III.

(and this is of course an important empirical question, one that has to be answered for each thing individually), then treating them as intrinsic goods in all of these situations will in fact be the pleasure-maximizing thing to do.

Perhaps surprisingly, there could also be a reason to think of as intrinsic goods even some things which have no instrumental value at all. It's possible that, even if an object we value is neither intrinsically nor instrumentally valuable, *the act of valuing it* is instrumentally valuable. It is an interesting and important fact about human beings that we are normally happiest *in the pursuit* of a goal. Finally achieving it does also normally bring us a fair amount of happiness, but it usually pales in comparison to the happiness we find in the mere anticipation of its achievement and in the process of working toward it. Simply having some goal to which we can dedicate our time and effort can greatly increase our happiness, quite apart from any benefits of actually achieving it. Thus, even if there happened to be no intrinsically valuable goals we could work toward, it seems that we would still have reason to adopt some ends, even if we had to do it at random (by manipulating our future selves into thinking certain things were intrinsically valuable, or by plugging ourselves into the experience machine). Of course, in the world as it is, there is no lack of objectively worthy goals. There are political, social, environmental, economic, medical, and personal problems to solve. We have no need to invent for ourselves frivolous goals, or preserve certain frivolous goals we already have, just to ensure ourselves the pleasure of working toward *something*.

But there might be yet another kind of reason to take things that have neither intrinsic nor instrumental value to be intrinsically valuable. Sprigge points out that we may have reason to value certain things if valuing them is part of an overall way of life that has greater "felt value." He considers in particular our taking human embryos and fetuses to be intrinsically valuable. He writes,

> …the proper attitude to embryos and fœtuses, is to be determined not only by considering the effect on their feelings, if they have them, but by considering the felt value of the way of life in which we experience them as having a certain sort of value. Thus we must ask whether investing them with the value with which we invest humans once they have been born (and which it would certainly vastly impoverish our lives not to do) makes for a life world which it is literally better to live in than any alternative to which we might move. To pose that question here is not to answer it. If, however, we do answer it positively, we should be careful not to raise the question again too often, for thereby we would cease to live that form of life in which we have seen such value.[243]

Sprigge doesn't make clear what the reasons might be for life's feeling more valuable if one values the lives of embryos and fetuses in this way. Perhaps there is some reason that valuing these things, rather than others, adds more pleasure to our daily lives. I am somewhat skeptical that it would actually be pleasure-maximizing to value anything that does not itself tend to produce pleasure, given that doing so would cause it to compete for our attention and energy with things that *do* produce it. But we should certainly acknowledge the possibility that there could be this sort of hedonistic reason.

[243] Sprigge, 211.

In the end, it seems that we will probably be right to try to eliminate our valuing of things that have neither intrinsic nor instrumental hedonistic value. On the other hand, we ought to preserve and encourage our tendency to think of as intrinsically valuable certain things that have consistently important instrumental hedonistic value—things like relationships, knowledge, and works of art—since thinking of these things as intrinsically good seems to be the most effective way of benefiting from their instrumental value. But if hedonists ought to approve of thinking of these things as having intrinsic value, then hedonism is not at odds with these common values, as is often thought. Granted, it doesn't take them at "face value"—i.e., as indicative of these things' actually *being* intrinsically valuable—but it nevertheless tells us to adopt exactly these attitudes to these things, and so the fact that we have these attitudes can hardly be a decisive strike against hedonism.

V. Human pleasure vs. animal pleasure

We now turn, however, to a subject on which many people's intuitions will *not* agree with the implications of hedonism, or rather, with the implications of the specific version of hedonistic utilitarianism I have developed in this dissertation. If instantiations of the phenomenal quality of pleasantness are the only thing that is intrinsically valuable, it seems quite possible that the intrinsic value of the life of a particular animal might be greater than the intrinsic value of the life of a particular human. A very happy pig, for instance, would seem on this theory to be living a more intrinsically valuable life than an only barely happy human being. This is at odds with many people's intuitions, and it's even at odds with the claim of that great defender of hedonistic utilitarianism John Stuart Mill, who advised us that "[i]t is better to be a human being dissatisfied, than a pig satisfied."[244] In this section, I will consider some possible hedonistic utilitarian justifications for Mill's claim, but I will ultimately conclude that someone who holds to my view about the nature of intrinsic goodness must reject it. I will argue that this implication of my view shouldn't count against it, however, given that there is reason to believe that our intuitions about this matter are distorted by a bias that is generally recognized not to reflect a moral difference.

Let's consider first Mill's own defense of the preferability of being a human being, even if one enjoys less pleasure than a pig. As is well known, Mill appealed to a difference between "higher" and "lower" pleasures. For Mill, less pleasure could be better, if it was of a particularly high quality, as he thought the pleasures "of the intellect, of the feelings and imagination, and of the moral sentiments" were.[245] Mill claimed that "higher" pleasures were unavailable to a pig, and that it was for this reason that human beings were superior to animals when it came to the intrinsic quality of their lives.

Mill's argument for the superiority of some pleasures over others was quite weak, however. To prove that some pleasures were "higher" and others "lower," Mill

[244] Mill, 252.
[245] Ibid., 250.

appealed to the fact that those who were capable of both types of pleasure chose the higher ones. He argued that those persons who did *not* devote themselves to intellectual pursuits or any of the other pleasures that were uniquely available to humans did not do so only because they had not developed the requisite *capacity*. Developing the capacity to enjoy the higher pleasures takes time and effort, and those who have not had the necessary training simply cannot know how superior the higher pleasures actually are.

What Mill did not discuss is whether those who devote themselves primarily to the "higher" pleasures may do so not because they are capable of both higher and lower pleasures and choose the former, but because they are incapable of enjoying the lower pleasures to the same extent as other people.[246] Someone who has had drilled into him his entire life the importance of scholarship is understandably going to have a difficult time feeling content with a life that consists of manual labor during the day and carousing in a bar at night. This may not be because the scholar's pleasures have a superior intrinsic quality but simply because the scholar can't help desiring to exercise his mental faculties. It seems doubtful that we have any entirely unbiased judges who, though capable of taking equal pleasure in intellectual and "mindless" pursuits, would always unreservedly choose the intellectual ones.

Even if this particular argument for the superiority of some pleasures fails, however, a hedonist might try to justify morally relevant differences among pleasures

[246] Sidgwick makes this point as well (148).

simply by appealing to our intuition that they exist, or by arguing that morally relevant qualitative differences can be found in the phenomenology of the pleasures themselves. Qualitative hedonistic theories have in fact been defended by contemporary philosophers: notably, Edwards and Crisp. However, someone who holds my view about the nature of intrinsic goodness can't help himself to qualitative hedonism. On my view, intrinsic goodness just is the phenomenal quality of pleasantness (this is what gives my view its epistemological advantages), and this means that a certain amount of pleasantness experienced by a pig has to be just as intrinsically good as the same amount of pleasantness experienced by a human being. No difference between pleasures that is not a difference in pleasantness can increase their intrinsic goodness because no other difference is a difference in *goodness*.

I don't think this is a conclusion to be ashamed of or to try to avoid, however. While it clashes with many people's intuitions, I believe there are good reasons to distrust our intuitions in this case. If we look at humanity's history of defining the boundaries of the moral community, we see that we have a decided tendency to privilege those who are like us, in race, sex, class, religion, and any number of other ways. This tendency has likely been somewhat beneficial to our survival. It is important, if we are going to recognize obligations to others and thus sacrifice some of our own interests from time to time, that those others generally reciprocate our sacrifices. And we seem to be able to be most confident of the reciprocation of others when they are very similar to us in culture and mentality. (Not to mention that altruism toward those most like us—our close relatives—directly promotes the survival of the genetic code that they share with us, including, perhaps, a gene that encourages altruism towards those similar to us.) But the mere fact that such bias was very likely instrumental in our species' survival, and thus in our own current presence on the scene, does not mean that it should be taken to indicate an objective fact about the relative values of races, sexes, cultures, or even species. In the early twenty-first century, much of the world seems to have eliminated from its official rhetoric any moral discrimination among races and sexes of human beings, but we still maintain that human happiness is of far more worth than that of non-humans. Perhaps the time has come to concede that this sort of discrimination cannot be justified any more than the others.[247]

This doesn't mean that we have to believe that porcine lives are just as valuable as human lives *all things considered*. Human lives do have certain *instrumental* value that pigs' lives do not. Humans have all sorts of abilities for complex thought and coordinated action that allow them to solve problems that pigs can't. When one is in a real pickle, it's generally better to have a human being around than a pig. Apart from concerns about solving problems, other human beings tend to be particularly instrumentally valuable to us because they make good company. Hanging out with a marginally unhappy person is often more rewarding than hanging

[247] For a much more detailed presentation of this argument, see Peter Singer, *The Expanding Circle: Ethics and Sociobiology* (New York: Farrar, Straus, & Giroux, 1981). Or, for a briefer version, see his *Animal Liberation* (New York: New York Review, distributed by Random House, 1975), Ch. 1.

out with a jubilant pig (although I can’t say there aren’t some times when I might prefer the pig). But if we restrict ourselves to talking purely in terms of the present intrinsic value of a very happy pig’s life and an only marginally happy person’s life, I believe we ought to concede that the pig’s life is better.

Yet there is another possible consequence of the equal value of human and pig pleasure that could be even more counterintuitive to some people. This consequence stems from the fact that a pig is likely a minor form of utility monster—that is, a pig likely requires somewhat less than a human in the way of resources in order to produce the same amount of pleasure. We might wonder whether this means that my version of hedonistic utilitarianism requires us to raise pigs rather than human beings.

The calculations of resource-to-pleasure ratios that would have to be done to determine whether this is actually the case are too involved to pursue here. One important factor to keep in mind when trying to decide the question, however, is a human’s ability actually to create or exploit resources that other animals could not profit from on their own. Though each human being may require more energy to sustain in a happy state, he may make up for his greater needs by his ability to produce food, fuel, etc. At the same time, I don’t think we can easily reject the possibility that, while it may be optimal for the total balance of pleasure over pain in the world to keep a certain number of problem-solving humans around, the optimal ratio of humans to members of other species may be much lower than the current one. While this may seem a little counterintuitive to many people, I doubt many people will be horribly repulsed by it, and it seems like the sort of idea that might come to seem very reasonable, once we get used to thinking of human beings as integrated parts of the natural world. In addition, as people in developed countries are becoming more and more aware of the finiteness of natural resources, they are realizing that we do have an interest in developing less expensive tastes, that perhaps we do have something to learn from the simple needs and pleasures of other animals. If we continue to feel the growing pressure of finite resources, it will likely make more and more sense that the only thing that can make one pleasure any better than another is its instrumental usefulness.

For these reasons, and because it seems likely that our original intuition about the superiority of human lives stems from a misleading bias against those different from us, I think that the intrinsic equality of human and animal pleasures implied by my view should not serve as a reason to reject the view but should instead be embraced.

VI. Conclusion

In this chapter, we have seen that our intuitions are generally not as contrary to hedonism as is often thought. We have seen that hedonism can justify our belief that there are both fitting and unfitting things in which to take pleasure and pain. We have also seen that hedonism justifies our negative feelings about living hooked up to the experience machine, first of all because of the need that we have to interact with the world on a regular basis in order to maximize our future happiness and that of others, and second of all because many of the things that advocates of the experience machine objection claim it shows are intrinsically valuable—things like relationships and knowledge—are things that there are hedonistic reasons for us to *think of* as intrinsically valuable, just because that's the best way of promoting their instrumental value.

On the other hand, we've seen that there is at least one intuition that some people are likely to have that my version of hedonistic utilitarianism cannot justify: the feeling that human pleasure is intrinsically more valuable than non-human animals' pleasure. But this feeling can plausibly be explained by a bias for those who are like us, the same bias that has led to past discrimination that we now recognize as unjustifiable. Thus this intuition does not seem like a good reason to reject a theory that is capable of grounding so many of our other moral intuitions in epistemically accessible, judgment-independent facts.

CONCLUSION

I hope to have made clear in this dissertation the importance of answering the central metaphysical and epistemological questions posed by moral realism: “What makes it the case that our concept of goodness objectively applies to certain things in the world?” and “How can we know to which things it objectively applies?” And I hope to have shown that there is at least one plausible way of answering these questions. Determining whether it is the most plausible way will depend on much further research, including research on other metaphysical and epistemological puzzles: the place of phenomenology in the universe, the possibility of knowledge of the external world, and even the nature of causation, which may affect our view of the link between normative phenomenology and motivation. I am hopeful that my core proposal—that the descriptive and the normative overlap in the directly epistemically accessible area of phenomenal experience—will seem increasingly logical as we begin to see the radical paradigm shifts necessary to answer these other important philosophical questions.

In the meantime, whether or not we believe that the goodness of pleasure and the badness of pain are the most *basic* moral facts, and whether or not we believe that their goodness and badness are judgment-independent, most of us believe that they are nevertheless very important, and this means that we can cooperate in our efforts to improve the balance of pleasure over pain in the world. There’s a lot to be done.

BIBLIOGRAPHY

Aristotle. *Nicomachean Ethics*.

Attfield, Robin. *A Theory of Value and Obligation*. London: Croom Helm, 1987.

Aydede, Murat. "An Analysis of Pleasure Vis-à-Vis Pain." *Philosophy and Phenomenological Research* 61, No. 3 (November 2000): 537-70.

Aydede, Murat, ed. *Pain: New Essays on Its Nature and the Methodology of Its Study.* Cambridge, MA: MIT Press, 2005.

Ayer, A. J. *Language, Truth and Logic*, 2nd ed. London: Gollancz, 1946.

Benacerraf, Paul. "Mathematical Truth." *Journal of Philosophy* 70, No. 19 (1973): 661-79.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. 1789.

Bernstein, M. *On Moral Considerability: An Essay on Who Really Matters*. Oxford: Oxford University Press, 1998.

Blackburn, Simon. *Spreading the Word*. Oxford: Clarendon Press, 1984.

Blackburn, Simon. "Errors and the Phenomenology of Value." In *Essays in quasi-realism*. Oxford: Oxford University Press, 1993. Pp. 149-65.

Blackburn, Simon. *Ruling Passions*. Oxford: Oxford University Press, 1998.

Boyd, Richard. "How to Be a Moral Realist." In *Essays on Moral Realism*. Edited by Geoffrey Sayre-McCord. Ithaca, NY: Cornell University Press, 1988. Pp. 181-228.

Brand, Paul, and Philip Yancey. *Pain: The Gift Nobody Wants*. New York: Harper Collins, 1993.

Brandt, Richard B. "Toward a Credible Form of Utilitarianism." In *Morality and the Language of Conduct*. Edited by Hector-Neri Castañeda and George Nakhnikian. Detroit: Wayne State University Press, 1963. Pp. 107-43.

Brandt, Richard B. *A Theory of the Good and the Right*. Oxford: Clarendon Press, 1979.

Brandt, Richard B. "The Science of Man and Wide Reflective Equilibrium." *Ethics* 100, No. 2 (Jan. 1990): 259-78.

Brandt, Richard B. *Morality, Utilitarianism, and Rights*. New York: Cambridge University Press, 1992.

Brink, David. *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press, 1989.

Brink, David. "Realism, Naturalism, and Moral Semantics." *Social Philosophy & Policy* 18, No. 2 (Summer 2001): 154-76.

Broad, C. D. *Five Types of Ethical Theory*. London: Routledge & Kegan Paul, 1930.

Buck, R. *Human Motivation and Emotion*. Chichester, UK: John Wiley & Sons, Inc., 1976.

Cabanac, Michel. "Optimisation du comportement par la minimisation du déplaisir dans un espace sensoriel à deux dimensions." *Comptes rendus des séances de l'Académie des sciences* 300, No. III (Paris, 1985): 607-10.

Cabanac, Michel. "Money versus pain: experimental study of a conflict in humans." *Journal of the Experimental Analysis of Behavior* 46 (1986): 37-44.

Cabanac, Michel. "La maximisation du plaisir, réponse à un conflit de motivations." *Comptes rendus des séances de l'Académie des sciences* 309, No. III (Paris, 1989): 397-402.

Cabanac, Michel. "Pleasure: the Common Currency." *Journal of Theoretical Biology* 155, No. 2 (1992): 173-200.

Cabanac, Michel. *La quête du plaisir: Etude sur le conflit des motivations*. Montréal: Liber, 1995.

Cabanac, Michel. "On the Origin of Consciousness, a Postulate and its Corollary." *Neuroscience and Biobehavioral Reviews* 20 (1996): 33-40.

Cabanac, Michel, and C. Ferber. “Pleasure and preference in a two-dimensional sensory space.” *Appetite* 8 (1987): 15-28.

Cabanac, Michel, and J. LeBlanc. “Physiological conflict in humans: fatigue vs. cold discomfort.” *American Journal of Physiology* 244, No. 5 (May 1983): R621–R628.

Carson, T. *Value and the Good Life*. Notre Dame, Ind.: University of Notre Dame Press, 2000.

Cicero. *De Finibus Bonorum et Malorum.*

Clark, Austen. “Painfulness is Not a Quale.” In Aydede 2005. Pp. 177-97.

Coghill, Robert C. “Pain: Making the Private Experience Public.” In Aydede 2005.

Coghill, Robert C., John G. McHaffie, and Ye-Fen Yen. “Neural Correlates of Interindividual Differences in the Subjective Experience of Pain.” *Proceedings of the National Academy of Sciences of the United States of America* 100, No. 14 (2003): 8538-42.

Crisp, Roger. “Hedonism Reconsidered.” *Philosophy and Phenomenological Research* 73, No. 3 (November 2006): 619-42.

Crisp, Roger. *Reasons and the Good*. Oxford: Oxford University Press, 2006.

Damasio, Antonio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace & Company, 1999.

Damasio, Antonio. *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. New York: Harcourt Brace & Company, 2003.

Damasio, Antonio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. B. Ponto, J. Parvizi, and R. D. Hichwa. “Subcortical and Cortical Brain Activity During the Feeling of Self-Generated Emotions.” *Nature Neuroscience* 3 (2000): 1049-56.

Darwall, Stephen. “Self-Interest and Self-Concern.” *Social Philosophy and Policy* 14 (1997): 158-78.

Darwall, Stephen, Allan Gibbard, and Peter Railton, eds. *Moral Discourse and Practice: Some Philosophical Approaches*. Oxford: Oxford University Press, 1997.

Diener, Ed, and Eunkook M. Suh. “National Differences in Subjective Well-being.” In *Well-Being: The Foundations of Hedonic Psychology*. Edited by Daniel Kahneman, Ed Diener, and Norbert Schwarz. New York: Russell Sage Foundation, 1999. Pp. 434-50.

Diogenes Laertius. *Lives of Eminent Philosophers*. Edited and trans. by R. Hicks. Rev. by H. Long. Cambridge, Mass.: Harvard University Press, 1972.

Dworkin, Ronald. “Objectivity and Truth: You’d Better Believe It.” *Philosophy and Public Affairs* 25, No. 2 (Spring 1996): 87-139.

Edwards, Rem B. *Pleasures and Pains: A Theory of Qualitative Hedonism*. Ithaca: Cornell University Press, 1979.

Eisenberger, N. I., and M. D. Lieberman. “Why Rejection Hurts: A Common Neural Alarm System for Physical and Social Pain.” *Trends in Cognitive Sciences* 8 (2004): 294-300.

Enoch, David. “An Argument for Robust Metanormative Realism.” Ph.D. diss., New York University, 2003.

Enoch, David. “Why Idealize?” *Ethics* 115, no. 4 (July 2005): 759-87.

Ewing, A. C. *The Definition of Good*. London: Routledge & Kegan Paul, 1948.

Ewing, A. C. *Ethics*. London: The Macmillan Company, 1953.

Feldman, Fred. *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*. Cambridge: Cambridge University Press, 1997.

Feldman, Fred. *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford: Clarendon Press, 2004.

Field, Hartry. *Science Without Numbers*. Princeton: Princeton University Press, 1980.

Field, Hartry. *Realism, Mathematics, and Modality*. Oxford: Blackwell, 1989.

Field, Hartry. "Which Undecidable Mathematical Sentences Have Determinate Truth Values?" In *Truth in Mathematics*. Edited by H. Garth Dales and Gianluigi Oliveri. Oxford: Clarendon Press, 1998.

Fine, Kit. "The Question of Realism." *Philosopher's Imprint* 1, No. 1 (2001): 1-30.

Finnis, John. *Natural Law and Natural Rights*. Oxford: Clarendon Press, 1980.

Finnis, John. *Fundamentals of Ethics*. Washington, D.C.; Georgetown University Press, 1983.

Firth, Roderick. "Ethical Absolutism and the Ideal Observer." *Philosophy and Phenomenological Research* 12, No. 3 (March 1952): 317-45.

Foltz, E. L., and L. E. White. "Pain 'Relief' by Frontal Cingulotomy." *Journal of Neurosurgery* 19 (1962): 89-100.

Foot, Philippa. "Moral Beliefs." *Proceedings of the Aristotelian Society* 59 (1958-9): 83-104.

Foot, Philippa. "Utilitarianism and the Virtues." *Mind* 94, No. 374 (April 1985): 196-209.

Fuchs, Alan E. "The Production of Pleasure by Stimulation of the Brain: An Alleged Conflict Between Science and Philosophy." *Philosophy and Phenomenological Research* 36, No. 4 (June 1976): 494-505.

Garcia, J., P. S. Lasiter, F. Bermudez-Rattoni, and D. A. Deems. "A general theory of aversive learning." *Annals of the New York Academy of Science* 443 (1985): 8-21.

Gibbard, Allan. "Utilitarianism and Human Rights." *Social Philosophy and Policy* Vol. I, Issue 2, Human Rights. Edited by Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul. Oxford: Basil Blackwell, 1984. Pp. 92-102.

Gibbard, Allan. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, Mass.: Harvard University Press, 1990.

Gibbard, Allan. *Thinking How to Live*. Cambridge, Mass.: Harvard University Press, 2003.

Gigerenzer, Gerd. *Gut Feelings: The Intelligence of the Unconscious*. New York: Viking, 2007.

Gigerenzer, Gerd, P. M. Todd, and the ABC Research Group. *Simple Heuristics That Make Us Smart*. New York: Oxford University Press, 1999.

Glynn, Ian. *An Anatomy of Thought: The Origin and Machinery of the Mind*. New York: Oxford University Press, 1999.

Gosling, J. C. B. *Pleasure and Desire: The Case for Hedonism Reviewed*. Oxford: Oxford University Press, 1969.

Gray, John. "Indirect Utility and Fundamental Rights." *Social Philosophy and Policy* Vol. I, Issue 2, Human Rights. Edited by Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul. Oxford: Basil Blackwell, 1984. Pp. 73-91.

Griffin, James. *Well-being: Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press, 1986.

Gustafson, Don. "Categorizing Pain." In Aydede 2005.

Gybels, J. M., and W. H. Sweet. *Neurosurgical Treatment of Persistent Pain*. Basel: Karger, 1989.

Hall, Richard J. "Are Pains Necessarily Unpleasant?" *Philosophy and Phenomenological Research* 49, No. 4 (June 1989): 643-59.

Hare, R. M. *Freedom and Reason*. Oxford: Clarendon Press, 1963.

Hare, R. M. "Nothing Matters." In *Applications of Moral Philosophy*. London: Macmillan, 1972. Pp. 32-47.

Hare, R. M. "Ethical theory and utilitarianism." In *Contemporary British Philosophy IV*. Edited by H. D. Lewis. London: Allen and Unwin, 1976. Pp. 113-131. Reprinted in *Utilitarianism and Beyond*. Edited by Amartya Sen and Bernard Williams. Cambridge: Cambridge University Press, 1982. Pp. 23-38.

Hare, R. M. *Moral Thinking: Its Level, Method, and Point*. Oxford: Clarendon Press, 1981.

Hare, R. M. “Utility and Rights: Comment on David Lyons’s Essay.” In *Ethics, Economics and the Law*, *Nomos* 24. Edited by J. Roland Pennock and John W. Chapman. New York: New York University Press, 1982. Pp. 148-57.

Harsanyi, John C. “Rule Utilitarianism and Decision Theory.” *Erkenntnis* 11 (1977): 25-53.

Hobbes, Thomas. *Human Nature: or the Fundamental Elements of Policy*. 1650.

Hodgson, D. H. *Consequences of Utilitarianism: A Study in Normative Ethics and Legal Theory*. Oxford: Clarendon Press, 1967.

Hogarth, R. M., and N. Karelaia. “Ignoring information in binary choice with continuous variables: When is less ‘more’?” *Journal of Mathematical Psychology* 49 (2005): 115-24.

Hogarth, R. M., and N. Karelaia. “Simple models for multi-attribute choice with many alternatives: When it does and does not pay to face tradeoffs with binary attributes.” *Management Science* 51 (2005): 1860-72.

Hogarth, R. M., and N. Karelaia. “Regions of rationality: Maps for bounded agents.” *Decision Analysis* 3 (2006): 124-44.

Honderich, Ted, ed. *Morality and Objectivity: A Tribute to J. L. Mackie*. New York: Routledge, 1985.

Horgan, Terence, and Mark Timmons. “New-Wave Moral Realism Meets Moral Twin Earth.” *Journal of Philosophical Research* 16 (1991): 447-65.

Horgan, Terence, and Mark Timmons. “Troubles on Moral Twin Earth: Moral Queerness Revived.” *Synthese* 92 (1992): 221-60.

Horgan, Terence, and Mark Timmons. “Moral Phenomenology and Moral Theory.” *Philosophical Issues* 15, Normativity (2005): 56-77.

Hume, David. *A Treatise of Human Nature*. 1740.

Hume, David. *An Enquiry concerning the Principles of Morals.* 1751.

Jackson, Frank. *From Metaphysics to Ethics*. Oxford: Clarendon Press, 1998.

Jackson, Frank. “Cognitivism, a priori deduction, and Moore.” *Ethics* 113, No. 3 (April 2003): 557-75.

Jacobson, Hilla. “The Evaluative Structure of Pain.” Work in progress, 2008.

James, William. *What Is an Emotion?* 1884.

Jollimore, Troy. “Meaningless Happiness and Meaningful Suffering.” *Southern Journal of Philosophy* 42, No. 3 (Fall 2004): 333-47.

Kagan, Shelly. “The Limits of Well-Being.” In *The Good Life and the Human Good.* Edited by Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul. Cambridge: Cambridge University Press, 1992. Pp. 169-89.

Kahneman, Daniel and Carol Varey. “Notes on the psychology of utility.” In *Interpersonal Comparisons of Well-being*. Edited by Jon Elster and John E. Roemer. Cambridge: Cambridge University Press, 1991. Pp. 127-86.

Katsikopoulos, K. and L. Martignon. “Naïve heuristics for paired comparisons: Some results on their relative accuracy.” *Journal of Mathematical Psychology* 50 (2006): 488-94.

Katz, Leonard. “Hedonism as Metaphysics of Mind and Value.” Ph.D. diss., Princeton University, 1986.

Kawall, Jason. “The Experience Machine and Mental State Theories of Well-being.” *Journal of Value Inquiry* 33 (1999): 381-7.

Kirchin, Simon. “Ethical Phenomenology and Metaethics.” *Ethical Theory and Moral Practice* 6 (2003): 241-64.

Kitcher, Philip. *The Nature of Mathematical Knowledge*. Oxford: Oxford University Press, 1983.

Korsgaard, Christine M. “Skepticism about Practical Reason.” *The Journal of Philosophy* 83, No. 1 (January 1986): 5-25.

Korsgaard, Christine M. “The Sources of Normativity.” In *The Tanner Lectures on Human Values*, Vol. 15. Edited by Grethe B. Peterson. Salt Lake City: University of Utah Press, 1994. Pp. 19-112. Excerpted in Darwall, Gibbard, and Railton. Pp. 389-406.

Korsgaard, Christine M. “Two Distinctions in Goodness.” In her *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press, 1996. Pp. 249-74.

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

Kymlicka, Will. *Contemporary Political Philosophy*. Oxford: Clarendon Press, 1990.

Lewis, David. “Dispositional Theories of Value.” *Proceedings of the Aristotelian Society* 63, Supplementary Volume (1989): 113-37.

Locke, John. *An Essay Concerning Human Understanding*.

Liotti, M., and J. Panksepp. “Imaging Human Emotions and Affective Feelings: Implications for Biological Psychiatry.” In *Textbook of Biological Psychiatry*. Edited by Jaak Panksepp. New York: Wiley, 2004. Pp. 33-74.

Lyons, David. *The Forms and Limits of Utilitarianism*. Oxford: Clarendon Press, 1965.

Lyons, David. “Utility as a Possible Ground of Rights.” *Noûs* 14, No. 1, 1980 A.P.A. Western Division Meetings (March 1980): 17-28.

Lyons, David. “Utility and Rights.” In *Ethics, Economics, and the Law*, *Nomos* 24. Edited by J. Roland Pennock and John W. Chapman. New York: New York University Press, 1982. Pp. 107-38.

Mackie, J. L. *Ethics: Inventing Right and Wrong*. London: Penguin, 1977.

MacLean, P. D. *The Triune Brain in Evolution*. New York: Plenum, 1990.

Maddy, Penelope. *Realism in Mathematics*. Oxford: Clarendon Press, 1990.

Maddy, Penelope. *Naturalism in Mathematics*. Oxford: Clarendon Press, 1997.

Mandelbaum, Maurice. *The Phenomenology of Moral Experience*. Glencoe, Ill.: Free Press, 1955.

Martignon, L., and U. Hoffrage. “Fast, frugal and fit: Lexicographic heuristics for paired comparison.” *Theory and Decision* 52 (2002): 29-71.

McDowell, John. “Projection and Truth in Ethics.” In Darwall, Gibbard, and Railton. Pp. 215-25.

McDowell, John. “Values and Secondary Qualities.” In Darwall, Gibbard, and Railton. Pp. 201-13.

McFarland, D. J., and R. M. Sibly. “The behavioral final common path.” *Philosophical Transactions of the Royal Society* (Series B), 270 (1975): 265-93.

McLelland, J. L., D. E. Rumelhart, and G. E. Hinton. “The Appeal of Parallel Distributed Processing.” In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1. Edited by D. E. Rumelhart, J. L. McLelland, and the PDP Research Group. Cambridge, Mass.: MIT Press, 1986. Pp. 3-44.

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

Mill, John Stuart. *Utilitarianism*. 1863. Reprinted in *The Six Great Humanistic Essays of John Stuart Mill*. Edited by Albert William Levi. New York: Washington Square Press, 1963. Pp. 241-308.

Millgram, Elijah. *Practical Induction*. Cambridge, Mass.: Harvard University Press, 1997.

Millgram, Elijah. “What’s the use of utility?” *Philosophy and Public Affairs* 29, No. 2 (Spring 2000): 113-36.

Monro, D. H. “Utilitarianism and the Individual.” *Canadian Journal of Philosophy* 5, Suppl. (1979): 47-62.

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

Moore, G. E. “Is Goodness a Quality?” In *Philosophical Papers*. London: Allen & Unwin, 1959. Pp. 89-101.

Morgan, M. M., M. M. Heinricher, and H. L. Fields. “Inhibition and Facilitation of Different Nocifensor Reflexes by Spatially Remote Noxious Stimuli.” *Journal of Neuropsychology* 72 (1994): 1152-60.

Nagel, Thomas. *The Possibility of Altruism*. Princeton: Princeton University Press, 1970.

Nagel, Thomas. "The Limits of Objectivity." Tanner Lecture on Human Values. Brasenose College, Oxford University, May 4, 11, and 18, 1979. Available from http://www.tannerlectures.utah.edu/lectures/documents/nagel80.pdf.

Nagel, Thomas. *The View from Nowhere*. New York: Oxford University Press, 1986.

Nagel, Thomas. *The Last Word*. New York: Oxford University Press, 1997.

Narveson, Jan. "Utilitarianism, Group Actions, and Coordination or, Must the Utilitarian be a Buridan's Ass?" *Noûs* 10, No. 2, Symposium on Utilitarianism (May 1976): 173-94.

Nathan, P. W. "Nervous System." In *The Oxford Companion to the Mind*. Edited by R. L. Gregory. Oxford: Oxford University Press, 1987.

Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.

Nozick, Robert. *The Examined Life: Philosophical Meditations*. New York: Simon and Schuster, 1989.

Panksepp, Jaak. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. London: Oxford University Press, 1998.

Panksepp, Jaak. "Social Pain, Support, and Empathy." In Aydede.

Panksepp, Jaak, B. H. Herman, T. Villberg, P. Bishop, and F. G. DeEskinazi. "Endogenous Opioids and Social Behavior." *Neuroscience and Biobehavioral Reviews* 4 (1980): 473-87.

Panksepp, Jaak, S. M. Siviy, and L. A. Normansell. "Brain Opioids and Social Emotions." In *The Psychobiology of Attachment and Separation*. Edited by M. Reite and T. Fields. New York: Academic Press, 1985. Pp. 3-49.

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

Price, D. D., A. Von Der Gruen, J. Miller, A. Rafii, and C. A. Price. "Psychophysical Analysis of Morphine Analgesia." *Pain* 22 (1985): 261-69.

Railton, Peter. “Alienation, Consequentialism, and the Demands of Morality.” *Philosophy and Public Affairs* 13, No. 2 (Spring 1984): 134-71.

Railton, Peter. “Moral Realism.” *Philosophical Review* 95 (April 1986): 163-207.

Railton, Peter. “Facts and Values.” *Philosophical Topics* 14 (Fall 1986): 5-31.

Railton, Peter. “Naturalism and Prescriptivity.” *Social Philosophy and Policy* 7, No. 1 (Autumn 1989): 151-74.

Rawls, John. “Outline of a Decision Procedure for Ethics.” *The Philosophical Review* 60, no. 2 (April 1951): 177-97.

Rawls, John. “Two Concepts of Rules.” *Philosophical Review* 64 (1955): 3-32.

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

Regan, Donald H. “How to be a Moorean.” *Ethics* 113, No. 3 (April 2003): 651-77.

Ross, W. D. *Foundations of Ethics*. Oxford: Clarendon Press, 1939.

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

Schachter, Stanley, and Jerome Singer. “Cognitive, Social, and Physiological Determinants of Emotional State.” *Psychological Review* 69 (1962): 379-99.

Sem-Jacobsen, C. W. *Depth-Electrographic Stimulation of the Human Brain and Behavior*. Springfield, Ill.: Charles C. Thomas, 1968.

Shafer-Landau, Russ. *Moral Realism: A Defence*. Oxford: Clarendon Press, 2003.

Shapiro, Stewart. *Philosophy of Mathematics: Structure and Ontology*. New York: Oxford University Press, 1997.

Sidgwick, Henry. *The Methods of Ethics*, 7th ed. Chicago: University of Chicago Press, 1907.

Silverstein, Harry S. “Utilitarianism and Group Coordination.” *Noûs* 13, No. 3 (Sept. 1979): 335-60.

Silverstein, Matthew. “In Defense of Happiness: A Response to the Experience Machine.” *Social Theory and Practice: An International and Interdisciplinary Journal of Social Philosophy* 26, No. 2 (Summer 2000): 279-300.

Singer, Peter. *Animal Liberation*. New York: New York Review, distributed by Random House, 1975.

Singer, Peter. *The Expanding Circle: Ethics and Sociobiology*. New York: Farrar, Straus, & Giroux, 1981.

Smart, J. J. C. “An Outline of a System of Utilitarian Ethics.” In *Utilitarianism: for and against*. By J. J. C. Smart and Bernard Williams. Cambridge: Cambridge University Press, 1973. Pp. 1-74.

Smith, Michael. “Should We Believe in Emotivism?” In *Fact, Science and Morality: Essays on A. J. Ayer's* Language, Truth and Logic. Edited by Graham Macdonald and Crispin Wright. London: Basil Blackwell, 1986.

Smith, Michael. “Reason and Desire.” *Proceedings of the Aristotelian Society* 88 (1987-88): 243-56.

Smith, Michael. “Dispositional Theories of Value.” *Supplement to the Proceedings of the Aristotelian Society* 63 (1989): 89-111.

Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1994.

Smith, Michael. “Response-Dependence Without Reduction.” *European Review of Philosophy* 3, Special Issue on Response-Dependence. Edited by Roberto Casati and Christine Tappolet (1998): 85-108.

Smith, Michael. “Does the Evaluative Supervene on the Natural?” In *Well-Being and Morality: Essays in Honour of James Griffin*. Edited by Roger Crisp and Brad Hooker. Oxford: Oxford University Press, 2000. Pp. 91-114.

Smith, Michael. “Moral Realism.” In *Blackwell Guide to Ethical Theory*. Edited by Hugh LaFollette. Oxford: Blackwell, 2000. Pp. 15-37.

Smith, Michael. “Exploring the Implications of the Dispositional Theory of Value.” *Philosophical Issues: Realism and Relativism* 12 (2002): 329-47.

Sobel, D. “Varieties of Hedonism.” *Journal of Social Philosophy* 33 (2002): 240-56.

Sprigge, T. L. S. *The Rational Foundations of Ethics*. London and New York: Routledge & Kegan Paul, 1988.

Street, Sharon. "Evolution and the Nature of Reasons." Ph.D. diss., Harvard University, 2003.

Street, Sharon. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127 (2006): 109-66.

Street, Sharon. "Constructivism About Reasons." In *Oxford Studies in Metaphysics*, Vol. 3. Edited by Russ Shafer-Landau. New York: Oxford University Press, forthcoming.

Sturgeon, Nicholas. "Moral Explanations." In *Essays on Moral Realism*. Edited by Geoffrey Sayre-McCord. Ithaca, NY: Cornell University Press, 1988. Pp. 229-55.

Sumner, L. W. *Welfare, Happiness, and Ethics*. Oxford: Clarendon Press, 1996.

Tännsjö, Torbjörn. *Hedonistic Utilitarianism*. Edinburgh: Edinburgh University Press, 1998.

Thomson, Garrett. *Needs*. London: Routledge & Kegan Paul, 1987.

Timmons, Mark. *Morality Without Foundations*. Oxford: Oxford University Press, 1999.

Toulmin, S. E. *An Examination of the Place of Reason in Ethics*. London: Cambridge University Press, 1950.

Urmson, J. O. "The Interpretation of the Moral Philosophy of J. S. Mill." *The Philosophical Quarterly* 3, No. 10 (1953): 33-9.

Wall, P. D. *Pain: The Science of Suffering*. New York: Columbia University Press, 2000.

Weiskrantz, L. *Consciousness Lost and Found*. Oxford: Oxford University Press, 1997.

Weiskrantz, L., E. K. Warrington, M. D. Sanders, and J. Marshall. “Visual capacity in the hemianopic field following a restricted occipital ablation.” *Brain* 97 (1974): 709-28.

Wiggins, David. “A Sensible Subjectivism.” In Darwall, Gibbard, and Railton. Pp. 227-44.

Williams, Bernard. “A Critique of Utilitarianism.” In *Utilitarianism: for and against*. By J. J. C. Smart and Bernard Williams. Cambridge: Cambridge University Press, 1973. Pp. 75-150.

Williams, Bernard. “Internal Reasons and the Obscurity of Blame.” In *Making sense of humanity and other philosophical papers*. Cambridge: Cambridge University Press, 1995. Pp. 35-45.

Williams, Bernard. “Internal and External Reasons.” In Darwall, Gibbard, and Railton. Pp. 363-71.

Williams, K. D. *Ostracism: The Power of Silence*. New York: Guilford Press, 2001.

Wright, Crispin. “Moral Values, Projection and Secondary Qualities.” *Proceedings of the Aristotelian Society* 62, Suppl. vol. (1988): 1-26.