University of Cincinnati

Date: 4/29/2011

I, Daniel F. Hartner, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Philosophy.

It is entitled:
Toward a Genuinely Natural Ethical Naturalism

Student's name: Daniel F. Hartner

This work and its defense approved by:

Committee chair: Valerie Hardcastle, PhD
Committee member: Vanessa Carbonell, PhD
Committee member: John Bickle, PhD
Committee member: Lawrence Jost, PhD
Toward a Genuinely Natural Ethical Naturalism

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Department of Philosophy

of the College of Arts and Sciences

by

Daniel F. Hartner

M.A. University of Cincinnati

May 2011

Committee Chair: Valerie Gray Hardcastle, Ph.D.
Abstract

Naturalism is an ambiguous philosophical term that refers to one of two general philosophical theses: the metaphysical thesis that all facts, including moral facts, are natural facts, and the epistemological thesis that the methods of philosophical inquiry, including moral inquiry, are continuous with those of the empirical sciences. The metaphysical thesis has long been the default form of naturalism in ethics. The thesis of this dissertation is that, first, despite the convention, it the epistemological thesis—which I call “methodological naturalism”—that is genuinely naturalistic; and second, that existing forms of epistemological naturalism in contemporary ethics have largely failed to develop genuine continuity between moral inquiry and scientific inquiry and hence are not genuine forms of methodological naturalism.

The argument proceeds in two parts. First, I examine a growing trend in moral philosophy toward the use of scientific data, especially from cognitive science, psychology, and psychiatry, for resolving longstanding philosophical disputes about the nature of morality and moral agency. This empirical strategy is widely regarded as naturalistic. I argue that many of these approaches fail to satisfy a fundamental requirement on genuine naturalism, namely the development of continuity with scientific inquiry, either because they wrongly take the metaphysical thesis rather than the methodological thesis as primary, or because despite rightly privileging the latter they nevertheless seek to resolve traditional philosophical disputes that are founded on an outmoded psychological framework, namely the framework of commonsense folk psychology. The traditional philosophical disputes in question are those that jointly constitute what Michael Smith (1994) calls the “Moral Problem,” the central organizing problem in metaethics. I focus in particular on the dispute surrounding the so-called Humean theory of moral motivation, according to which an agent’s being motivated to act morally requires not just a belief about what is morally right but also a desire to act in accordance with that belief.

The second part of this dissertation applies naturalistic methodology to the Humeanism dispute. I synthesize data ranging from cellular-level neuroscience to cognitive science, social neuroscience, and neuroeconomics to sketch a philosophically informed scientific model of moral cognition and motivation. I argue that this model is in tension with the commonsense psychological framework presupposed by the Moral Problem and by the Humeanism dispute in particular. Central to this argument is the idea that cellular level data on decision-making, which is too often regarded as philosophically irrelevant, can be brought to bear directly on the kind of data from cognitive science that is frequently cited in moral philosophy. The model of moral judgment, decision, and motivation that results, I argue, is a much more promising starting point for philosophical inquiry into moral agency than the framework of commonsense folk psychology. Genuinely naturalistic approaches to ethics must aim to develop real continuity between philosophical inquiry into morality and scientific inquiry into human agency. This requires that we use the best available scientific data to vet the empirical assumptions at work in traditional philosophical theories rather than to merely use that data to adjudicate between them.
Acknowledgments

I would like to thank my dissertation director, Valerie Hardcastle, who graciously supervised this project, and my dissertation committee, Larry Jost, Vanessa Carbonell, and John Bickle, for their many helpful comments, criticisms, and discussions of the arguments in this dissertation. I am especially indebted to John, my advisor, for his patience, encouragement, and, above all, for his example. Most of what I know about neurophilosophy I learned by following John’s lead.

I would also like to thank the faculty and graduate students in the Department of Philosophy at the University of Cincinnati for their generosity, support, and guidance, and perhaps most of all for providing an ideal environment for interdisciplinary philosophy. I couldn’t have found a better department anywhere for this research project. I have benefitted greatly from conversations with many of the students and faculty, especially Tom Polger, Larry Jost, Vanessa Carbonell, and Aaron Kostko.

I would also like to thank my family, especially my parents and siblings, for their limitless support, which came in many forms, and without which I would not have survived graduate school.

I wrote much of this dissertation while funded by a Graduate Student Research Fellowship provided by the University Research Council at the University of Cincinnati, and by a Taft Dissertation Fellowship provided by the Charles Phelps Taft Research Center at the University of Cincinnati. I am very grateful for the generous support of these organizations.
Table of Contents

Abstract ii
Acknowledgments iv
Table of Contents v
List of Figures vi

1. Introduction: Methodological Naturalism & the Moral Problem 1
   §1. Preliminaries 1
   §2. The Fundamental Dilemma of Metaethics 10
   §3. A Dogma of Ethical Naturalism 15
   §4. Naturalism from Moore’s Open Question to Smith’s Moral Problem 26
   §5. Conceptual Analysis & Quasi-Naturalism 35
   §6. An Outline of the Master Argument by Chapter 36

2. Moral Motivation and Philosophical Psychology 44
   §1. Preliminaries 44
   §2. The Humeanism Dispute 45
   §3. Belief/Desire Realism 48
   §4. Humeanism & the Moral Problem 54
   §5. Toward a Dilemma for Metaethics 62
   §6. Of Commonsense Theories of Moral Motivation 66

3. Functionalist Moral Psychology & the Humeanism Dispute 91
   §1. Preliminaries 91
   §2. Functionalism: From Mind to Moral Psychology 92
   §3. Canberra Functionalism in Moral Psychology 101
   §4. Naturalism vs. the Canberra Plan 104
   §5. Canberra-Style Armchair Psychology 126
   §6. More Shoddy Psychology as Conceptual Analysis 135
   §7. Some Quasi-Naturalistic Approaches 139
   §8. Toward the Science of Moral Motivation 142

4. Operational Definitions in Philosophical Psychology 144
   §1. Preliminaries 144
   §2. The Absence of Operational Definitions in Moral Psychology 145
   §3. Operational Definitions for the Humeanism Dispute 161

5. Humeanism and the Neuroscience of Decision 167
   §1. Preliminaries 167
   §2. Valuation and the Common Currency Hypothesis 168
1. Introduction: Methodological Naturalism and the “Moral Problem”

§1. Preliminaries

The goal of this dissertation is to motivate a new way of thinking about the philosophical position known as “methodological naturalism” in ethics. Naturalism—most broadly the view that reality consists in nothing over and above the natural world of rocks, trees, brains, cells, molecules, and atoms, which we study in the empirical or natural sciences—is trendy in philosophy these days, especially in the philosophy of mind and ethics. Rapid advances in the neural and cognitive sciences in the past few decades have yielded a vast collection of interesting cases, experiments, and data to use in support of philosophical intuitions about the relations between mind and body, biology and morality.

In one sense this is promising. The traditional view of philosophy as largely autonomous and independent from other forms of inquiry into human agency is beginning to fall out of favor. Some philosophers have even begun collecting their own data (see “experimental philosophy”) and contributing to “moral psychology,” a relatively nascent research program situated at the confluence of ethics and psychology (see, for example, recent work by Casebeer, Greene, Nichols, and Stich).

Still, somewhat paradoxically for a self-described naturalist, I am going to argue that this growing movement in its current form is ultimately the wrong kind of empirical turn for moral philosophy. I will try to explain in some detail why the strand of naturalism in which philosophers apply new neuroscience to old philosophical problems—the kind of quasi-

---

1 By “moral psychology” here I mean to refer exclusively to the relatively new, interdisciplinary study of moral cognition exemplified by these authors and not the relatively old subfield of traditional metaethics that is sometimes given the same name.
naturalism, as I will call it, that is fast becoming the most popular approach of any kind to doing moral philosophy—is problematic.² It is time we overhaul some traditional philosophical issues by using data from the empirical study of the human nervous system as the point of departure. This, I will argue, is the way to a more genuine or complete kind of naturalism. To make the case concrete, I will focus my argument on one of the more popular topics in metaethics and moral psychology at present: the problem of moral motivation.

Naturalists are right to recognize the relevance of the empirical study of the brain to philosophical questions about morality. But they concede the crucial part of the naturalistic project when they force scientific data to square with traditional philosophical positions and problems. The conceptual foundations for many of these traditional problems were developed long before philosophers had access to the empirical data that now heavily constrains philosophical frameworks and theories. In comparison to metaethics, moral psychology is embryonic. Much of the data that has immediate relevance to philosophical questions about moral agency, and moral motivation in particular, has only been generated in the last couple of decades or even the last few years. But that data is already beginning to reveal that several traditional philosophical questions are likely outmoded.

The difficult empirical work required to adequately defend this claim will occupy the second half of this dissertation. The positive contribution of this dissertation is thus a philosophically informed synthesis of independent lines of scientific research, ranging from

² Many of the philosophers discussed in this essay would likely object to the label “quasi-naturalism.” I intend the term to reflect my view that these projects are incomplete forms of naturalism, not theirs. I think the term “naturalism” is overused. It should be reserved for approaches seriously committed to developing real continuity between philosophy and the sciences. That requires attending, from the outset, to best available data immediately relevant to the philosophical inquiry, in this case, data concerning the nature of judgment and motivation in the cognitive and neural sciences. To my mind there are very few projects in contemporary ethics that meet the criteria.
cellular neuroscience to social neuroscience and neuroeconomics, on moral decision and motivation. The empirical model of moral decision that emerges from this synthesis of the best available scientific data on decision and motivation is at once intended as both a direct challenge to the empirical assumptions about the nature of cognition at work in contemporary metaethics and a new, more empirically adequate point of departure for philosophical inquiry into the nature of morality.

Specifically, I will argue that the core disputes about moral motivation in metaethics, and the dispute about the Humean theory of moral motivation in particular, presuppose either implicitly or explicitly both the existence of folk psychological mental states and their causal relevance to empirically adequate explanations of moral motivation and moral behavior. But the updated empirical model shows that the present science of decision and motivation—which gets directly connected to disputes about moral judgment, decision, and motivation via research concerned with the contribution of emotion to both economic and moral and social decision—does not presuppose such folk psychological states, does not invoke them, and likely will not provide an adequate foundation for the claim that such states “supervene” on (or must somehow be preserved by) the immediate causal mechanisms of decision and motivation. Our best available explanations for decision and motivation are revealing causal mechanisms for these phenomena that are unlikely to lend themselves to accurate interpretation in the coarse vocabulary of commonsense psychology.

---

3 This thesis has much in common with eliminativism, roughly the view that commonsense mental states do not exist. Though I am sympathetic to that view, my goal here is not to defend eliminativism in this broad sense. Rather, my aim is only to defend the much narrower claim that the mental states assumed in the disputes situated at the core of metaethics do not figure in the relevant ways into our best scientific explanations of moral decision and motivation.
Perhaps most importantly, I will emphasize that this argument is much less grand in its ambitions and less reaching in its conclusions than it first seems. It is far less reaching and considerably more moderate than the practice of continuing to import an old empirical framework containing untested empirical claims about the nature of cognition and motivation into contemporary philosophical debate. That contemporary neural and cognitive science will or must preserve both the structure and causal relevance of a commonsense theory of cognition that is hundreds of years old should be a significant result that emerges from careful empirical inquiry—as profound as the claim that that future physics will or must preserve our 18th Century intuitions about the nature of the physical world or the laws that govern it—rather than an unremarkable assumption upon which to build an entire philosophical field.

This is a project concerned with the current state of the various branches of the neural sciences. It is not an argument about the development of philosophical inquiry into human moral agency, or about the history of the relationship between philosophy and the sciences of the mind/brain. Until very recently it made sense to view a mostly a priori armchair philosophy as just one among several different points of entry for developing theories of human cognition and behavior. There is good reason to think that this has changed. The empirical research projects in psychology, cognitive science, and neuroscience that were at one time tentative and only tangentially relevant to the interests of philosophers are now generating models of cognition and behavior that likely have immediate implications for the empirical adequacy of many traditional philosophical debates about moral agency, such as those concerning the mechanisms of moral judgment, moral cognition, motivation and the like.

We should recognize that many of these research programs were originally inspired by big philosophical questions about morality. But it doesn’t follow from this observation about the
origin and development of the neural and cognitive sciences that traditional philosophical questions must remain empirically relevant in the face of rapid advances in those sciences. Nor does it follow that traditional conceptual frameworks must continue to serve as the basis for future research projects. As a science of moral agency develops, we should expect that most of the conceptual questions would need to be revised or even discarded, and that new philosophical questions will arise out of existing lines of empirical research.

Yet philosophers who investigate any aspect of human psychology, moral philosophers especially, persist in the idea that philosophical methods and empirical data have something close to an equal role in generating constraints on the development of philosophical questions and theories. Here, for example, are two contemporary naturalists describing their own research projects in philosophical psychology:

The conceptual territory is well developed, and it is now reasonable to think that whether an externalist or an individualistic interpretation of a given psychological phenomenon is warranted is an empirical issue. Nevertheless, there are classically philosophical tasks that can contribute to empirical investigation of externalism. One obvious thing we can do is to draw conceptual distinctions that then can be used for framing psychological hypotheses. Another less obvious thing we can do is to ask, perhaps in a preliminary manner, whether externalism would be important if it turned out to be true (Sneddon 2008: 394)

Philosophers have generally tried to establish the link between emotion and moral judgment by armchair reflection. I think philosophical analysis is a good way to make progress on the conceptual question: can one possess a moral concept without having certain sentiments… As a starting place, I want to focus on a more tractable question to consider: do our ordinary moral concepts (the ones we deploy in token thoughts most frequently) have an emotional component? This is essentially an empirical question… Empirical questions can be addressed using philosophical methods (philosophical intuitions can be treated as data), but laboratory studies are useful as well. In the spirit of promiscuity, I propose to intermingle empirical and philosophical results (Prinz 2006: 30).
In both cases naturalism is construed as an intermingling of traditional philosophical projects and empirical research. More specifically, both take the philosophical contribution to involve drawing conceptual distinctions useful for formulating and testing empirical hypotheses.

It will be no part of my argument to deny that traditional philosophical methods can still make contributions to empirical inquiry and in particular to the development of empirical hypotheses. Purely conceptual questions can and do inspire and aid empirical research programs and theories. But the crucial point is that these programs must begin by acknowledging the empirical constraints set by existing data. And that means understanding the relevant research before committing to broad conceptual frameworks.

The question for Sneddon is why we should bother to ask whether externalism would be important if true before we have any idea what externalism (or its chief opposition) would really amount to as an empirical theory of human cognition. It isn’t just obvious that the assumptions underlying the conceptual questions are supported by facts about the human nervous system. Yet, despite conceding that the question is ultimately empirical, Sneddon reverts to a traditional conceptual framework. On what basis can we really hope to draw conceptual distinctions that might be useful to empirical inquiry if we haven’t first looked to the data that will ultimately constrain how we frame psychological hypotheses about the boundaries of our cognitive mechanisms? The methods we must now use to develop relevant hypotheses about the minds and brains of real human agents may turn out to bear little resemblance to traditional philosophical methods.

---

4 This seems to be an increasingly popular move in philosophical psychology these days. In later chapters I will provide several examples in which philosophers offer disclaimers about the legitimacy of empirical assumptions and then carry on with the issue anyway.
In Prinz’s case one problem is that conceptual questions about the relationship between moral judgment and emotion are already thoroughly constrained by the meanings of these terms in the relevant sciences of cognition and behavior. More specifically, it is difficult to imagine how we might make philosophical progress on the relationship between ordinary moral concepts and concepts of emotion given that the latter gain their meaning from particular research programs. Prinz is right to recognize that the question whether moral concepts have an emotional component is an empirical one. But, as I will argue in later chapters, the philosophical concept of emotion *simpliciter* is giving way to fine-grained distinctions between kinds of emotional processing. These distinctions are generally made on the basis of the particular neural regions and pathways that are involved in the processing of a stimulus or in the execution of some cognitive task. Our traditional concept of emotion is just too vague to figure into or parallel the kinds of detailed explanatory hypotheses that the neural sciences are now generating. Consequently, depending on how we construe the term we can get very different answers to Prinz’s question about whether our ordinary moral concepts have an emotional component.

To put the point in vaguely Carnapian terms, the rate at which the research is developing has made philosophical questions like this nearly meaningless outside of a particular research framework. To be sure, conceptual questions remain for philosophers. But those questions will arise naturally out of inquiry built on solid empirical foundations. And this is what ought to motivate naturalism.

Owen Flanagan made this point two decades ago in his “principle of minimal psychological realism”:

Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for creatures like us (1991: 32).
Moral philosophers have too often failed to abide by this principle. Darwall, Gibbard and Railton (1997) lamented this fact in their overview of the last century of ethical inquiry, and Doris & Stich (2005) have recently tried to address it: “Regarding the assessment of Darwall and colleagues, we couldn’t agree more: Far too many moral philosophers have been content to invent the psychology or anthropology on which their theories depend, advancing or disputing empirical claims with little concern for empirical evidence” (2005: 114). Despite increasing awareness of the problem, however, few have undertaken the kind of empirical work needed to correct it. Most forms of contemporary ethical naturalism are just not really all that naturalistic, and have failed to appreciate just how far the neural sciences have come.

In an effort to facilitate some of that change, I will argue that naturalism in ethics, as in any branch of philosophical inquiry, is properly a commitment to a particular methodology – one that takes seriously the importance of the scientific method and the data it produces, seriously enough in fact to insist upon careful empirical investigation even when we might be able to get away with casual reflection upon our own experiences or even selective attention to supportive data in its place. To adopt methodological naturalism is to insist that empirical claims, especially those that bear on important philosophical issues, be examined or held up for comparison with the existing data before they are assumed and used as a basis for broad conceptual questions, generalizations and theories about the way we critters work.

In the rest of this introductory chapter I lay the groundwork for this argument. In §2 I discuss in very general terms the development of some of the foundational issues in ethics and metaethics. I focus primarily on moral properties since this sets up my own view of what ethical naturalism ought to amount to. Even so, the section is primarily intended to motivate the problem in intuitive terms. It therefore traces the admittedly oversimplified picture of moral
philosophy that tends to serve as a starting point for further discussion in the works of other philosophers, particularly Michael Smith (1994). It is not intended to be a comprehensive treatment of ethical theories, and those already familiar with the foundational issues in metaethics can probably skip ahead without much loss of continuity. In §3 I explain naturalism as a general philosophical stance, distinguish methodological naturalism from metaphysical naturalism, and explain why I think genuinely naturalistic approaches in ethics ought to focus on the former rather than the latter.

In §4 I introduce an influential framework for thinking about the relationship between the foundational problems in contemporary metaethics, namely Michael Smith’s (1994) “moral problem.” Smith calls the moral problem the central organizing problem in metaethics because it helps to explain why metaethical disputes engender so much disagreement and why those disagreements take the form that they do, with philosophers clustering into fairly predictable camps. I think that one of the most direct routes to appreciating the need for a genuine methodological naturalism is to show that this central organizing framework is fundamentally flawed and to show how forging more direct links between moral inquiry and the science of human agency can help to fix it.

In §5 I draw a contrast in general terms between methodological naturalism and one primary source of opposition, a philosophical project known as the “Canberra Plan.” The Canberra Plan is an approach to philosophical methodology that blends a particular kind of a priori conceptual analysis with an affinity for naturalism made manifest by its interest in locating philosophical concepts within the empirical sciences. The Canberra Planners—Frank Jackson, Philip Pettit, and Michael Smith, in particular—have applied their particular approach to some of the philosophical issues constitutive of the moral problem, including the problem of moral
motivation. One of my main lines of argument for methodological naturalism will involve showing why I think the Canberra Plan will ultimately fail to make headway on the problem of moral motivation and so too in metaethics and moral psychology more generally. Like the many self-described naturalists engaged in quasi-naturalistic philosophical projects that I address in this essay, the Planners too have failed to appreciate the need for a more thoroughly empirically informed approach to moral philosophy. Finally, in §6 I outline the master argument for this essay and the contribution that the subsequent chapters are intended to make toward it.

§2. The Fundamental Dilemma of Metaethics

Ethical theory and applied ethics—what are generally referred to together as normative ethics—are the familiar branches of philosophy that seek to develop general moral principles, to specify the right-making features of our actions, and to investigate questions about the application of moral values, our standards of right and wrong, to particular circumstances. For example, “Is it morally wrong to kill sentient creatures because we enjoy eating them, and why?” are questions in normative ethics. Metaethics by contrast is the philosophical investigation of ethics itself.\(^5\) Whereas normative ethics deals with first-order questions, or questions of right and wrong, metaethics deals with second-order questions, that is, questions about those first-order questions. In normative ethics we ask whether some action, thought, or behavior is right or wrong; in metaethics we ask what rightness and wrongness are, or to what such moral vocabulary actually refers, or whether it refers to anything at all. Metaethics is not concerned

\(^5\) Of course, the boundary between normative ethics and applied ethics is not always so clear, and I don’t mean to imply that metaethics has a monopoly on philosophical questions about morality. These paragraphs are intended only to lay out in very general terms some of the differences between the subfields of moral philosophy.
with what has moral value, but rather with what is moral value. It is the study of ethics itself. For this reason it is sometimes called analytic ethics.

Metaethical questions arise naturally in moral discourse. Because our ordinary talk about what is right and wrong refers to moral properties, we are inclined to ask philosophical questions about those properties: What is it exactly that makes something morally wrong? Where in the world do we find the wrong-making feature or property of that action? The popular argument that genetic engineering is wrong because it is unnatural, for example, rests on the premise that moral rightness is a property somehow tied up with the unaltered (presumably by humans) “natural” order of the universe or the evolutionary development of species. It relies on claims about the kind of thing right-making properties must be.

Many of the interesting philosophical questions that we ask about ethical properties or features of the world arise because claims about ethical properties are purportedly objective (Darwall 1998; Smith 1994). If I assume that your claim about the moral wrongness of genetic engineering is intended to apply only to you and to no one else, then much of the incentive for asking difficult philosophical questions about your claim is probably removed. Since your claim applies to you and no one else, determining what grounds it isn’t a particularly pressing matter. But when you claim that genetic engineering is wrong for everyone in virtue of its interfering with the natural order of the universe, an order to which we all must be answerable, I will probably start to wonder what kind of thing the natural order of the universe really is and how you came to know so much about it.

The fact that ethical claims seem to be the sort of things that derive their force from ethical properties leaves us wondering what such properties are like, whether they really exist,
and how we come to know about them. The trouble is that it is really hard to say exactly what kind of thing moral wrongness is or, for that matter, just where it is supposed to be located.

Hume pointed out this very problem in the *Treatise*. When we reflect on premeditated murder, we tend to regard the act as having the moral property of being vicious. Yet, upon further investigation of the events surrounding the murder, we are unable to “locate” (literally or figuratively) that particular property. Hume says:

> Take any action allow’d to be vicious: Wilful murder, for instance. Examine it in all lights, and see if you can find that matter of fact, or real existence, which you call vice. In which-ever way you take it, you find only certain passions, motives, volitions, and thoughts. There is no other matter of fact in the case. The vice entirely escapes you, as long as you consider the object (III.1.26, p. 301).

Stephen Darwall calls this “Hume’s Challenge” (1998: 22). There is a problem about locating (or perhaps specifying or picking out) the properties from which our ethical judgments get their purported objectivity. It arises naturally when we pause to ask why certain features of a situation or action make that action seem morally wrong to us. As Hume says, when we ask what it is about premeditated murder that makes it morally repugnant, we have no trouble locating the matters of fact, such as the murderer’s state of mind (disregard for human life, violation of laws and social norms, etc.), but we are at a loss when it comes to locating the property of being vicious.

We might try to solve the problem by appealing to some objective and universal truth whereby all situations and actions possessing particular features are always morally wrong. But this doesn’t really answer Hume’s challenge. For even if we postulate objective and universal moral truths, we are still left unable to say how we come to ascertain these universal truths, what counts as evidence for them, or how we came to establish such generalizations without circular appeal to the ethical properties that give rise to Hume’s challenge in the first place.
It seems clear that we do not come to know ethical properties in the same way that we come to know other physical features of the world, namely by the sensory experiences with which we explore the physical world. If we did, then Hume’s challenge would seem easy to answer, and we would have no trouble locating the ethical properties in question. If this is right, then it cannot be true that we come to establish universal ethical truths by empirical observation as we do with scientific generalizations.

So, short of postulating an additional human sense or intuition for picking up on ethical properties, it remains difficult to explain our interaction with ethical properties, and Hume’s challenge remains unanswered. Some philosophers have responded to this problem by postulating such a human sensory faculty or intuition. There is something intuitively unsatisfying about this solution since it is hard to imagine what this mysterious moral sense faculty is like and how it could be so different from our other sensory faculties. And we are still left to wonder exactly what it is that our ethical intuitions are picking up on.

This problem turns out to be closely related to another one of Hume’s well-known observations: the fact-value distinction. In the Treatise, Hume argues that facts and values are fundamentally different kinds. No statement about what *ought* to be the case can be validly derived from statements about what actually *is* the case. The mere fact that we can make meat products cheaper and more readily available through the use of large-scale factory farming operations does not allow us to validly derive the normative conclusion that we *ought* to.

---

6 Or so the traditional story goes. Some philosophers have better resources at their disposal for responding to Hume’s challenge. For example, Paul Churchland appeals to the facts about our connectionist neural architecture to explain our responses to moral situations. Still, it is probably fair to say that few such solutions have really caught on and so Hume’s challenge persists. But again I must emphasize that here I am simply outlining the basic issues in metaethics, largely following Michel Smith, rather than attempting a comprehensive overview of the field since its inception.
Facts seem to be much more directly connected to reality than values. They are fairly objective and reasonably stable across observers. This seems to be much less the case with values. Values seem to depend much more on the characteristics of individual observers, which is perhaps why they engender so much controversy. Yet previously I said that philosophical questions about ethical judgments seem to arise naturally because those judgments have objective purport. That you think murder is wrong for all people everywhere—that you take this to be a fact about the way the world is—leads me to ask metaphysical and epistemological questions, like “How do you know?” But defensible answers are hard to come by.

This is, on the stock story, the crux of metaethical inquiry. Moral discourse is plagued by a tension between objectivity and subjectivity, practicality and generality, facts and values. Moral claims whenever they are put forward seem to be the kind of thing about which we can be right and wrong, yet it is entirely unclear just what in the world there is for them to be right and wrong about. Moral claims are fundamentally practical: they bear directly on what we are to do under particular circumstances. But for this reason they seem to be expressions of desires rather than statements about general matters of fact or the way that the world actually is. For common sense suggests that it is our desires (or desire-like states) that really move us.

The objective purport of moral judgments implies that such judgments are the expression of beliefs about moral facts, while the practicality of moral judgments implies just the opposite, that is, that those judgments are not beliefs at all but rather the expressions of our desires. But how could anything be both the expression of a belief in an objective fact and the expression of a subjective desire? This problem is sometimes called the “fundamental dilemma” of metaethics (Darwall 1998) or, as we shall see momentarily, the “moral problem” (Smith 1994). The various
schools of thought in contemporary analytic ethics can in general be viewed as attempts to explain (or explain away) this tension.

Two basic strategies are plain: on the one hand, we can avoid the problem of specifying the referents—presumably ethical properties of some kind—of moral judgments by giving up their purported objectivity; and on the other hand, we can preserve the objective purport of moral judgments by facing down some very difficult questions about the mysterious ethical properties to which those judgments refer. This is at bottom the dispute between ethical realists and anti-realists (or nihilists). Ethical realists think there are objective ethical facts. Ethical nihilists think that there are no such facts.

Importantly, both strategies seem to yield unpalatable consequences. If moral judgments do not really appeal to objective facts, as the nihilist claims, then we are forced to explain why we talk and act as though they do. Alternatively, if moral judgments have objective purport, as the ethical realist claims, then we are faced with explaining what strange sort of fact could be capable of making moral judgments true. I will return to this problem shortly when I present Smith’s more formal version. First I need to say something about how ethical naturalism fits into this picture.

§3. A Dogma of Ethical Naturalism

Naturalism in the broad sense is a general philosophical stance that refers to one of two general ideas: either the metaphysical claim that nature is all that there is or that all facts are natural facts, or the epistemological (or methodological) idea that scientific methods set the
standard for philosophical inquiry. Thus the term “naturalism” is ambiguous and may be used to refer to one or both of these distinct theses. The distinction is important because I think it is methodological naturalism and not metaphysical naturalism that must ultimately restore progress in metaethics. On my view, genuine philosophical naturalism is a commitment to methodological naturalism; genuine ethical naturalism is the application of genuine philosophical naturalism to moral philosophy. But as things stand now this conflicts with the stock view of the relationship between philosophical and ethical naturalism.

On the stock view, metaphysical naturalism holds that the only things that exist are natural, i.e., the sort of things generally studied by empirical investigation or things that have properties that can be modeled mathematically – things like rocks, trees, chairs, cells and molecules (and perhaps also numbers, relations, and so on). Methodological naturalism holds that the model method for acquiring knowledge about the world—irrespective of one’s metaphysical views—is that of the empirical sciences. Within the context of ethical inquiry, naturalism generally refers to the view that the facts about ethics, whatever those are, are facts about the natural world. To distinguish this specific ethical view from the more general type of philosophical naturalism it is typically called ethical (or metaethical) naturalism.

Traditionally, then, ethical naturalists are metaphysical naturalists who think that there are ethical facts. They follow the objective purport of ethical discourse to its intuitive conclusion: our ethical discourse has objective purport because it is discourse about the way the world actually is. This means that ethical naturalism is generally regarded as a form of ethical realism by definition (cf. Darwall 1998: 28-29). Ethical realism (also called “moral realism” – I say that it is a “stance” rather than a position because there are many different and competing versions of naturalism and, perhaps surprisingly, they are not all in agreement on what makes a philosophical position natural.  

---

7 I say that it is a “stance” rather than a position because there are many different and competing versions of naturalism and, perhaps surprisingly, they are not all in agreement on what makes a philosophical position natural.
use these terms interchangeably) is the view that there are truth-makers for ethical propositions. Realists claim that our ethical judgments express propositions, which are capable of being true or false, and that there really are objective features of the world that determine those truth-values. For this reason, ethical realism is itself one form of cognitivism, which is the view that ethical sentences (e.g., “Murder is wrong”) express propositions and are therefore truth-apt (i.e., capable of being true or false).

Within ethics, then, naturalism is primarily a *metaphysical* position. The metaethical landscape as it is now carved up forces the ethical naturalist to commit to the ontological thesis that there really are moral facts in order to commit to the claim that those facts are somehow natural facts. That is why noncognitivism is technically never a form of *ethical* naturalism despite the fact that many noncognitivists are self-described *philosophical* naturalists. The noncognitivist claims that ethical sentences are not truth-apt because those sentences do not describe any objective feature of the world. Ethical facts cannot be natural facts because there are no ethical facts at all. So, as things stand now, ethical naturalism requires an advance commitment to moral realism, or the existence of ethical facts, and to a primary interest in the metaphysics of moral value.

I want to briefly challenge this idea here because the strand of methodological ethical naturalism that I develop in this essay is not concerned with establishing ontological conclusions or with making broad metaphysical assumptions about the nature of morality in advance. It is rather a methodological thesis about how best to study the nature of moral agency whatever the consequences for disputes about the metaphysical status of morality. We don’t need to know whether there really are moral facts in order to know how it would be best to go about studying morality. It seems wrong to think that one must make deep metaphysical commitments *before*
joining the naturalistic movement in ethics. Moreover, because the title “ethical naturalism” now connotes metaphysical baggage, it almost seems that philosophical naturalists, and especially those who are genuinely interested in developing continuity between philosophy and scientific methodology, ought to tend toward ethical nonnaturalism. I find this confusing. Ethical naturalism and philosophical naturalism ought to be united by something other than a claim to the same title. The term is overused.

It seems to me that the reason ethical naturalism has come to be regarded as a metaphysical thesis primarily is that we tend to assume that ontological commitments determine methodological commitments rather than vice versa. The assumption may in part be a relic of the historical development of moral inquiry and the naturalistic movement, which I’ll say more about in the next section. But perhaps some find the idea intuitive for the following reason. Once we have some idea about what the universe is made up of we will have a better idea of how to go about studying it. In ethics, then, once we come to the conclusion that all of the facts about morality are somehow just facts about the natural world, we will know that we should study moral facts by studying natural facts, and that may involve some empirical science. Conversely, science is concerned with uncovering facts about the world, and there is no point in applying the scientific method to moral inquiry if there are no moral facts to be uncovered in the first place. Consequently, it seems intuitive that methodological naturalism should follow on the heels of metaphysical naturalism. But I think this idea is really only intuitive to the extent that we accept a subtle dogma, one popular in moral philosophy, about the nature of science.

The problematic dogma in my view is that the mark of the scientific is a commitment to studying a particular ontology, some subset of the furniture of the natural world, rather than a commitment to a particular method of investigating the only reality that we know. More
specifically, most contemporary ethical naturalists seem to just assume that the purpose of studying morality by engaging with empirical data is primarily to learn about the nature of a particular class of entities or properties in the universe, namely moral properties, by engaging with scientific ontologies, i.e., the subset of the world’s furniture that the sciences are currently in the business of studying, or that the sciences presuppose. This is problematic because it is just as plausible that, for example, the study of morality is primarily about how the creatures who instantiate or otherwise interact with moral properties—whatever those properties are or whether there really are any at all—actually work. Even some moral realists have made use of this observation (e.g., Railton 1986a). And the plausibility of that idea suggests that we can try to carry out a careful investigation of morality via moral agency by adopting the same standards or methods that we would use to study any other human activity without saying or knowing much of anything at all about the metaphysics of morality in advance.

The idea that there are no moral properties with which human agents might interact should be a profound conclusion drawn from a long process of inquiry into moral cognition and behavior, not an initial assumption with which to begin. That is why I think moral philosophers may have missed one of the few points of consensus to emerge from the philosophy of science in the last century: the scientific enterprise is characterized much more by its method than by its subject matter (Gorham 2009: 53). This confusion continues to hamper naturalism in metaethics.

---

8 In fact I think some strands of noncognitivism, especially some strands of expressivism, have been motivated by this idea, even if not explicitly. While expressivism might be metaphysical in the sense that it is a view about moral facts or properties (namely that there aren’t any), it seems to be motivated by the idea that we can learn a lot about what kinds of things moral properties are or aren’t by thinking about what it is that moral agents do when they refer to them.

9 This is not to deny that the scientists have some tacit metaphysical commitments. My point is merely that those commitments are, despite the oft-heard objections about the dangers of romanticizing or idealizing science, not the primary mark of scientific inquiry and not as essential to the possibility of scientific inquiry as some philosophers would have us believe. I develop this argument in more detail below.
For those who doubt that naturalists concern themselves primarily with ontology, or treat science as a set of metaphysical claims, it is worth noting that some leading ethical naturalists commit to the metaphysical thesis explicitly. They observe that the sciences have serious ontological commitments and that to engage with those sciences is to accept those ontologies.

Consider first a legitimately naturalistic version of naturalism. Casebeer (2003), a scientifically-minded philosopher who actually argues that the methodological thesis must be primary, actually points out that in taking the methodological and epistemological assumptions of the natural sciences as standards for inquiry he must “avail myself of the ontologies postulated by the natural sciences during the course of this inquiry… with the fallibilistic view that the ontologies of our current sciences might be wrong” (9). Consequently, despite claiming to take the methodological constraint as primary, he acknowledges that his project presupposes ontological naturalism “to a certain extent,” and even that it could be upended by bad scientific ontologies.

Casebeer’s scientifically adept naturalism is metaphysically modest in comparison to the rest of the field. The vast majority of naturalistic moral philosophers begin with metaphysics, though they are generally less forthright about it. David Copp (2007) for example defends a version of ethical naturalism according to which moral properties are empirical properties (40). He distinguishes natural properties from nonnatural properties on the basis of epistemological access. A property is natural if and only if any synthetic proposition (i.e. a non-analytic proposition, or a proposition which is true in some way other than in virtue of the meanings of its terms) about its instantiation that can be known could only be known empirically, or through experience (39-40).
At first this sounds a lot like the general idea behind methodological naturalism. Like Casebeer, Copp says that naturalism is an epistemological commitment at bottom. And he also acknowledges that “the reason we are tempted to turn to science in explicating naturalism… is that we take science to be our most reliable source of empirical knowledge” (39). Despite that, Copp’s naturalism is different from methodological naturalism because it is primarily concerned with the nature of properties. He is concerned with the epistemology because of his prior ontological ideas about natural properties or the natural world:

On my proposal, ethical naturalism is the position that moral properties are empirical properties… The empirical conception of naturalism seems to me to answer to the fundamental intuition behind naturalism. The primitive intuition is that the natural world is the world around us, the world we are immersed in. The key problem is to explain the sense in which we are ‘immersed’ in it. My proposal is to explain this on the basis of the nature of our epistemic access to the world. The natural world is the world we know empirically (40).

The epistemological commitment expressed in the last sentence—one with which I might otherwise agree—is rooted in the prior metaphysical commitment, namely an “intuition” about the kind of thing the natural world is.

This metaphysical idea about what the natural world is also explains why he wants to distinguish naturalism from “scientism.” We are tempted to turn to naturalism, Copp says, because we believe science to be our most reliable source of empirical knowledge. But there are other sources of empirical knowledge, and so there is no good reason to tie naturalism to science. Naturalism says that natural properties are empirical, not that they must be properties that figure into scientific theory (39). He says:

If mental properties are epiphenomenal, for example, they presumably will play no role in the true scientific story of the world, yet for my purpose they ought to count as natural. Naturalism is not scientism, and here is the place to distinguish the two by insisting that the naturalist holds only that natural properties are empirical, not that they must be properties that figure in scientific theory (39).
The primary concern is with what natural (and moral) properties are. My objection is that contemporary ethical naturalists, including those who are relatively friendly to integrating scientific research in their own work, seem to think that the characteristic feature of scientific inquiry is its ontology rather than its methodology.

Most ethical naturalists are “naturalistic” in the sense that they are committed to metaphysical ideas about the nature of moral entities or properties. So when it comes to the role that the sciences must play in uncovering truths about the nature of those entities or properties, they ask how scientific ontologies match up with their preconceived ideas about moral ontologies. Where they don’t match, so much the worse for the sciences. We should avoid commitment to a science-centered naturalism because some properties, such as moral properties or mental properties, might not appear in scientific ontologies, and we don’t want to preclude them from being “natural” on that basis. This view of naturalism rests on the dogma that the sciences are distinguished by their ontologies rather than by their methods. And that assumption in turn gives way to the idea that ethical naturalism requires a prior commitment to the existence of moral facts.

For example, Copp’s concern is that there are some entities and properties—dollar bills, or the property of being a unit of currency or the like are the only other examples he gives—that don’t figure into scientific theory but that are still known empirically. There are empirical sources of knowledge of entities and properties other than the sciences. So if we happen to acquire that knowledge of those properties without doing science, we still want to be able to claim that those properties are natural. But even if that is true, it is irrelevant.

Even if such entities or properties aren’t actually found in any scientific theory right now, the pertinent point is that they could be. They are natural in the sense that they are open to the
best form of investigation we have for learning significant truths about the empirical world. In fact it is hard to imagine any other sense in which they could be natural. If they are natural because they are empirical, then they are natural in the sense that, like all empirical properties, they are amenable to observation. So they must also be amenable to systematic observation. Even those not particularly drawn toward “scientism” have noticed this. Here, for example, is Alan Gibbard: “Natural properties, we might say, are ones that could figure in empirical science, and if spooks and gods were recognizable, they might be subjects of empirical science” (2003: 99).

Thus naturalism is scientism in Copp’s sense of the term precisely because the scientific method is applicable to any empirically observable property, even when we don’t bother to actually apply it. This is why his point about epiphenomenalism is irrelevant. If mental properties turn out to be epiphenomenal then naturalism in any meaningful sense of the term will be dead in the water. But the relevant point is not that naturalism will have stalled because epiphenomenal properties aren’t scientific (i.e., they aren’t actually studied by scientists) but because epiphenomenal properties aren’t accessible via empirical investigation at all, and so we have no hope of learning about such properties by systematically observing the natural world.

Naturalism in the face of this finding would be dead not because epiphenomenal properties don’t figure into scientific theories but because they couldn’t. If they aren’t “scientific” it is because they aren’t empirical. Driving an artificial wedge between the empirical and the scientific won’t save naturalism (whatever that is) should epiphenomenalism turn out to be the true account of the mental. It just seems that way so long as we are willing to view
science as a warehouse containing the subset of the universe’s empirical properties that we have investigated systematically. But that is a misleading picture of science.\(^1\)

Copp’s distinction between the empirical and the scientific only appears defensible because it invokes some artificial examples of empirical knowledge that we might not care to challenge. But knowledge of street names and dollar bills are weak instances of empirical knowledge. These are instances of knowledge of social conventions or stipulations, much like dictionary definitions. Because such knowledge has little to do with inductive generalization, or the uniformity of nature to use a classic phrase, it seems to be a good example of knowledge that is empirical but non-scientific. But such knowledge, the acquisition of which requires nothing more than referencing a map or a posted sign, does not seem to be empirical in a meaningful sense. Maps and signs contain predicates the extension of which is one particular stretch of road in one particular neighborhood as designated by some property owner. Knowledge of facts about the street itself—its topography, geography, and so on—seems a clearer case of empirical knowledge, but it also seems much less obviously non-scientific.

The names or values of dollar bills is a similarly strange example because each token of a bill type is named according to its value and it makes absolutely no difference whether your five-dollar bill and mine are of the same or different types (e.g. made up of identical material, created

---

\(^1\) One might reply on Copp’s behalf that the metaphysical claim is supposed to follow from the epistemological claim, which is primary. More charitably, then, we should read Copp as claiming that ‘natural properties are empirical’ rather than as claiming that ‘natural properties are empirical.’ But this reply not only misses the point, it exemplifies the problem. To defend a thesis about what makes moral properties natural, even the thesis that they are natural because they are empirical or scientific, is to move away from the methodological project. Genuine continuity between moral philosophy and science means primarily applying, in practice, systematic empirical methodology to moral inquiry, not primarily defending a metaphysical thesis about the nature of properties. The distinctive characteristic of science—and thus of genuinely methodological naturalism—is its methodology and precisely not its (empirical/natural/scientific) ontology.
at the same place or time or by the same press, etc.) so long as the people working in a
government office somewhere agree to acknowledge that they have the same economic value.
We don’t apply the inductive method to knowledge of stipulations because such knowledge isn’t
really empirical and because it has little to do with nature, not because it isn’t scientific.

When we move away from these cases of social convention and stipulation to actual
knowledge of the external “natural” world, the interesting question is whether our empirical
knowledge is applicable to further cases. This is particularly true in contemporary metaethics. It
might seem a non-scientific empirical fact that agents tend to be motivated by their moral
judgments, but if we really want to know the extent to which that generalization holds we have
some careful and systematic observation to do, at least if we want the claim to be relatively
uncontroversial. Because any such empirical observation is amenable to testing via the
scientific method, the only legitimate forms of non-scientific empirical knowledge are those so
obvious that they perhaps require no such testing to be widely accepted as applicable to further
cases without controversy. But non-scientific in practice is not non-scientific in principle.

Science is a method of studying whatever phenomena we can observe or learn to observe
in the external world. For example, the scientific method will be carried out the same way
whether you are interested in reaction times in psychology, the healing power of prayer, or the
causal efficacy of mental states. Anything that is empirically observable or that can be defined
in such a way as to make it empirically observable can be studied using the scientific method.
It’s just that in many cases interesting results won’t be forthcoming. The problem with
phrenology isn’t that its faulty ontology precludes us from doing it. To the contrary, phrenology

---

11 I add the point about controversy because we are too often willing to treat controversial
generalizations about human agency as commonsensical. My point here is about the norms
governing empirical generalization and not about how those norms sometimes get ignored in
appeals to “common sense” in moral philosophy. I provide examples of the latter in Chapter 3.
remains a live option. We can still observe variations in the features of cranial topography and compare them to variations in neural and cognitive function even today. The relevant point is that if we do it right—if we are careful in our methodology—we will likely find no result worth reporting. Similarly, cellular neuroscience will not go the way of phrenology even if it turns out that cells are not real entities in the universe in some literal sense. Nor for that matter would psychology become a pseudoscience should it turn out that there are no such entities as minds in the traditional sense of the term. This is precisely why I began by taking issue with Casebeer’s disclaimer.\textsuperscript{12}

To sum up, the distinction between the empirical and the scientific is one without a difference. Viewing science as a subset of empirical properties has created a divide between philosophical naturalism and ethical naturalism. Because the question raised by naturalism is the question of the relationship between philosophy and the sciences, and because so many moral philosophers take an ontological view of the sciences, moral philosophy tends to work with an ontological view of naturalism. But naturalism, like science, is a methodological ideal incapable of molding itself to prior ontological commitments. This is the sense in which naturalism advocates \textit{continuity} between philosophy and science (e.g., Darwall et al. 1992). It seems to me that confusion about this point is one of the pillars of quasi-naturalistic ethical naturalism.

\textbf{§4. Naturalism from Moore’s Open Question to Smith’s “Moral Problem”}

The stock view of ethical naturalism, with its focus on the nature of ethical properties, has been long in the making. Since G.E. Moore introduced the “open question argument” in his

\textsuperscript{12} Some recent arguments against fMRI-based research employ this strand of reasoning, which seems to conflate scientific ontologies and scientific methods. I consider those arguments briefly in Chapter 5.
1903 *Principia Ethica* moral philosophers of a naturalistic persuasion have sought to show how they can define the predicate “good” in natural terms, the kinds of terms employed in the natural sciences, without committing a logical fallacy. Moore argued that any attempt to define goodness (moral or otherwise) in terms of some natural property (e.g., pleasantness, desirability, etc.) leaves open the question whether anything thought to have that property (e.g., telling the truth, ameliorating human suffering, etc.) is good. Presumably this question (e.g., “Is telling the truth good?”) is perfectly intelligible regardless of what natural property is offered as a definition of “good.” For this reason it seems that the open question argument rules out any attempt to define goodness in terms of any natural property and, consequently, also the possibility of naturalizing ethics.

Ethical naturalists replied with a variety of reasons to think that the open question argument is itself invalid. Consequently, few philosophers today accept Moore’s argument. Even so, most contemporary moral philosophers are willing to admit that Moore is onto something interesting, even if it isn’t clear exactly what that is (Darwall et al. 1992). The general, albeit vague, intuition seems to be that naturalistic definitions of moral concepts neglect some important feature of what it is for a thing to be moral (Lenman 2008).

Concrete proposals vary, though a collection of objections to naturalism (e.g. Gibbard 2003, Mackie 1977, Parfit 1997, among others) suggests at least one general possibility: that any naturalistic definition of moral concepts will fail to accommodate the motivating element of those concepts. That is, naturalistic conceptions of morality seem unable to explain how moral judgments guide actions or motivate moral agents to act accordingly. This, of course, is the normative component of our moral judgments, manifested in the “ought” of our moral language, which at least since Hume philosophers have taken to be a species fundamentally different and
underivable from descriptive claims.\textsuperscript{13} Thus two (at least remotely) related questions come to bear on the possibility of achieving a naturalized ethics: how can naturalistic definitions of the good both answer Moore’s open question and accommodate the putatively unique action-guiding feature of moral judgments?

This challenge to moral naturalism is thus linked to a metaethical debate about whether the reasons why an action is morally right or wrong are \textit{necessarily} reasons to act morally. The position that they are is called \textit{internalism}, since the reasons for taking a particular course of action are internal to the demands of morality. The denial of this position, called \textit{externalism}, claims instead that moral reasons are not necessarily reasons for acting. In fact there are at least three distinct though related debates in metaethics that employ the internalism/externalism terminology with, I think, unfortunate consequences for the literature in ethics. The formulation given above is sometimes called the \textit{morality/reasons internalism/externalism} debate (e.g., Darwall 1997) or simply \textit{reasons internalism/externalism} (e.g., Rosati 2006, Williams 1981), and is probably the foundational version of the debate. Hobbes and Mill, for example, were externalists about moral reasons, claiming that the considerations that make an action right or wrong are not themselves reason giving. Kant was perhaps the first to reject this view outright. For Kant, the moral law—his famous Categorical Imperative—is internal to moral agents: agents give themselves the moral law in their practical reasoning through free and rational deliberation (Darwall 1998, Sedgwick 2008).

\textsuperscript{13} This is Hume’s famous \textit{is-ought} problem, which is often confused with Moore’s naturalistic fallacy according to which it is logically impermissible to infer the proposition that X is good on the basis of propositions about X’s natural properties. The related, though distinct, is-ought problem is the observation that there is something fundamentally different about descriptive and prescriptive statements such that the latter cannot be validly derived from the former.
At least two other internalism/externalism debates are derivative upon the dispute about reasons internalism/externalism, one of which is, it seems to me, particularly germane to the naturalist’s project. To help settle the dispute between reasons internalists and externalists, for example, we might ask what makes a consideration a reason for acting. Or, to put it a bit differently, what is the relation between practical reasons and motivation? According to *judgment internalism* (hereafter just “internalism”), when an agent S judges (or believes) that she should take some action A then, *necessarily*, S has some motivation to do A.

Internalism, unsurprisingly, has been given a few different formulations, which vary in the degree of the motivational strength that they attribute to the agent. On one of its weakest and most common formulations, judgment internalism holds that moral judgments motivate necessarily though defeasibly. In other words, internalists claim that an agent cannot hold a moral belief without being motivated—at least to some degree—to act, even if that motivation is ultimately defeated by other considerations and therefore fails to produce the requisite action.\(^{14}\)

The denial of this view is *judgment externalism*. As it turns out, one’s position on this particular internalism/externalism debate tends to reveal a good deal about one’s commitments on broader metaethical questions, particularly concerning cognitivism, Humean psychology, and, I think, perhaps even naturalism too.

To see why this is so, consider the following purportedly inconsistent set of claims, which is due to Michael Smith (1994, p.12):

\[^{14}\text{One might hold a still weaker formulation according to which moral judgments motivate “for the most part”. The trouble with these especially watered-down formulations, it seems to me, is that we might start to wonder how many exceptions the view can admit before its necessity claim begins to seem uninformative. At some point the view starts to become indistinguishable from externalism. It seems to me we could raise this concern about any version of internalism that employs a defeasibility condition. But I won’t argue for the point in this essay. Instead I focus on the other half of the moral motivation literature, namely the dispute surrounding the Humean theory of motivation.}\]
1. Moral judgments of the form ‘It is right that I $\phi$’ express a subject’s beliefs about an objective matter of fact, a fact about what it is right for her to do.

2. If someone judges that it is right that she $\phi$s then, \textit{ceteris paribus}, she is motivated to $\phi$.

3. An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume’s terms, distinct existences.

At least roughly, we can say that (1) expresses cognitivism, (2) judgment internalism, and (3) Humean psychology. By “Humean psychology” (or just “Humeanism” here) I mean the standard interpretation of Hume’s position on human psychology according to which beliefs and desires are the two basic kinds of psychological states. Beliefs, on this view, are states that purport to represent the world as it actually is, while desires are states that represent the way we want the world to be. Beliefs, since they claim to represent the world as it is, must be true or false. This is not so for desires, which make no claim about the way the world in fact is. Desires, unlike beliefs, are not even subject to rational criticism on this view. The Humean theory of motivation in ethics is usually expressed rather simply as the idea that motivation requires, in addition to a moral belief, a conative state, such as a desire. This is the idea expressed in (3).

The inconsistency arises because (1), (2), and (3) together entail a necessary connection between moral beliefs and desires while (3) denies that very connection. Moral judgments express beliefs (“cognitivism”), which according to internalism are necessarily connected to motivation, and Humeanism tells us that it is desires that provide that connection. From this we
get the necessary connection between beliefs and desires, but it is this connection that Humeanism explicitly denies in claiming that beliefs and desires are distinct existences.\textsuperscript{15}

The idea is that moral philosophers are forced to choose which of these three intuitive ideas to jettison. If moral judgments express beliefs (cognitivism), and those beliefs necessarily motivate (internalism), then it cannot be the case that moral motivation requires, in addition to a belief, a desire (i.e., Humean psychology must be false). Alternatively, we can maintain Humeanism by rejecting either cognitivism or internalism. If motivation requires a desire to act in addition to a moral judgment, and moral judgments are necessarily motivating, then it must be false that moral judgments express beliefs, i.e., some form of non-cognitivism, such as expressivism, must be true. And, finally, we might preserve both Humeanism and cognitivism by rejecting internalism. Moral judgments express beliefs, motivation to take action requires a desire in addition to that belief, and so moral judgments themselves do not motivate necessarily. In short, then, non-cognitivists reject (1), externalists reject (2), and so-called anti-Humeans reject (3).\textsuperscript{16}

As I said previously, how ethicists should handle the tension between the objectivity and practicality of moral judgments, a tension made salient by the above inconsistent triad, is what Smith calls “the moral problem”. Ethicists set out to make sense of ordinary moral practice, but such practice is plagued by tension between the apparent objectivity of moral judgments and the

\textsuperscript{15} In any case this is how Smith, the originator of this particular formulation, explains the tension. It seems to me clearer to simply say that (1) and (2) together entail the falsity of (3). In any case the result—an inconsistency in maintaining these three claims simultaneously—is the same, so nothing of obvious importance rides on how we explain it.

\textsuperscript{16} There are also philosophers who attempt to accommodate all three propositions within this framework, such as Smith (1994), who claims that what has been overlooked is a distinction between motivating reasons and normative reasons, and Mackie (1977) who claims that cognitivism is true and that moral judgments are intrinsically motivating, but that all such judgments are false because there are no moral facts (error theory).
intuitive idea that such judgments are expressions of our desires (1994, 11-12). If Smith is right that this is the central organizing problem in contemporary metaethics, then it should now be clear why one’s position on the internalism/externalism debate tends to reveal their commitments on broader metaethical questions such as Humeanism and cognitivism. Now let’s return to naturalism.

What Moore’s open question argument was thought to show (though, again, few philosophers today still think the argument succeeded) is that “good”, the fundamental property in ethical inquiry, is simple or atomic (i.e., it admits of no further division into constituent parts), unanalyzable, and irreducible. One historical consequence of Moore’s argument was the ascension of noncognitivism in metaethics (Darwall et al. 1997). By interpreting the acceptance of a moral judgment as an attitude of categorical endorsement, noncognitivists, unlike naturalists, can avoid the problem of identifying the particular natural property that moral judgments are thought to pick out, and so too the naturalistic fallacy. In other words, one historical consequence of the open question argument was to show naturalistically inclined moral philosophers that their only viable option was to reject (1) in Smith’s framework. As most naturalists (excluding eliminativists) were inclined to take seriously both the distinction between beliefs and desires as two fundamental kinds of psychological states and the unique action-guiding feature of moral judgments without which moral inquiry loses some of its luster, it was cognitivism—the truth-aptness of ethical sentences—that they would have to reject. And by and large this is just what they did, at least until about the 1950s.

What came next of course were a variety of challenges to noncognitivism. Those challenges were each rooted in one of the many philosophical developments that characterized the period, ranging from Wittgenstein’s treatment of language to Quine’s rejection of the
analytic/synthetic distinction. Consequently, the naturalistically inclined went back to looking for alternative approaches to handling the fundamental moral problem without running afoul of the basic Moorean intuition that *something* about the concept of “goodness” makes it impervious to explication in natural terms, even if that feature can ultimately be accounted for in other ways (and hence the naturalistic fallacy avoided). For example, some invoked the distinction between analytic and synthetic property identities (e.g., Boyd 1988). So for any natural property N, it is an open question whether X, which has the property N, is good because “N” and “good” differ in meaning, and yet it doesn’t follow that “N” and “good” can’t refer to the same property (Darwall 1998). The situation, they argued, is analogous to that of “water” and “H₂O”, which differ in meaning but refer to the very same thing.

Yet on approaches like these, it becomes hard to see how the cognitive moral judgments that refer to natural-cum-moral properties can be necessarily motivating, as internalism demands. The ability to retain (1) seems to come at the cost of (2). And so these philosophers are obliged to reject internalism in favor of externalism since the natural properties in virtue of which moral judgments are made truth evaluable are, at least ostensibly, not intrinsically motivating. A moral agent, the “sensible knave”, might well believe that telling the truth promotes, say, social utility and simply not care. Presumably the same is true of desirability, pleasantness, collective happiness, or any other candidate natural property.

On pains of consistency, then, naturalists (who, for the reasons discussed above tended toward cognitivism) tend to adopt judgment externalism. Cases, real or hypothetical, in which agents hold moral beliefs and simply choose not to act on them, are frequently cited to make the view more plausible. Infamous psychopaths who have claimed to hold the belief that killing is wrong while doing it anyway, or generic “amoralists” whose moral judgments offer them no
incentive to take the requisite action, are adduced to illustrate the possibility of moral judgments completely divorced from motivation – a conceptually impossible state of affairs from the perspective of internalism.

What I think is important about this brief sketch of the development of metaethics is that it demonstrates just how deeply the moral problem is entrenched in ethical inquiry. And in so far as ethicists continue to take seriously the above framework as the central organizing problem in contemporary metaethics, it is hard to see where naturalism might go from here. The philosophical constraints on the development of metaethical theories are firmly established. At best, new data might help us decide which well-trodden philosophical path is to be preferred. Beyond that, the empirical investigation of human cognition and behavior, despite the remarkable rate at which it continues to develop, has little left to offer metaethics and moral philosophy. It should perhaps be unsurprising then that so many versions of naturalism employ the backwards, quasi-naturalistic strategy of forcing new data to fit old philosophical problems.

To put the point a bit dramatically, to reverse course is to risk breaking away from some entrenched philosophical assumptions hundreds of years in the making. Quasi-naturalism offers the best of both worlds: the appearance of an empirically respectable research project coupled with a healthy deference for philosophical tradition. I think this is a harmful compromise for the discipline that Aristotle called “practical philosophy”. I think we can pay our respects to the history of philosophy without engaging in pseudo- or quasi-naturalistic research programs. For that reason I want to try to resist this trend by chipping away at the above framework, and hence the central organizing problem in metaethics, beginning with the dispute surrounding proposition (3).
§5. Conceptual Analysis and Quasi-Naturalism

One of the paradigm cases of contemporary quasi-naturalism is the so-called “Canberra Plan.” The Canberra Plan is a research program that developed mostly out of collaboration at the Australian National University in Canberra over the last couple of decades. It is a movement most closely associated with the work of David Lewis, Frank Jackson, Philip Pettit, and Michael Smith (henceforth “JPS” when I refer to their collective work). Canberra Planners are largely united by their commitment to a particular philosophical method that Daniel Nolan (2009) has called the Canberra Plan “two-step.” The goal of the two-step is to provide a philosophical analysis of the central concepts at work in some domain, such as human psychology or morality.

In the first step we round up the platitudes concerning the concepts at work in the domain of interest. What exactly platitudes are is a matter of contention, but the general idea is that platitudes are basic truths about the domain of interest, say for example human mental states. These might be the ordinary beliefs we have about human mental states, the meanings implicit in our use of mental state terminology, descriptions of paradigm cases of mental states, etc. Once we have a set of platitudes about the nature of human psychology, then the theoretical role for the foundational ideas about human psychology is supposed to emerge.

We end up with a conjunction of sentences about human psychology that contains both “outsider” terms (O-terms) and “insider” theoretical terms (T-terms). O-terms are terms that get their meaning from outside the theory while T-terms are those that get their meaning from the role they play inside the theory. T-terms are then “Ramsey eliminated,” or replaced, by existentially quantified variables. Since, however, we have only eliminated the T-terms and not the things over which they range, we have to look to the world to determine what might play the role in the theory described by the terms contained in the platitudes. This is the second step in
the Canberra two-step. Generally it involves looking to the sciences to find what might realize the roles specified in the first step. Though it isn’t a necessary feature of the plan, most Canberra Planners are in fact physicalists. Most think that if in the second step we find no physical realizers then we will find no realizers at all. In part the method is motivated by the desire to find in the physical world things that are not obviously physical, such as, to take the same example, the human mind (Braddon-Mitchell and Nola 2009: 4-9).

The method is paradigmatically quasi-naturalistic because, as Braddon-Mitchell and Nola put it, it “sets an agenda of analysis combined with a penchant for the naturalistic, given the involvement of science” (9). The traditional conceptual analysis occupies the first step while the naturalistic move occupies the second. Consequently, the Canberra Plan, with its philosophy-first approach to analyzing the domain of interest, is among the major opponents of the kind of methodological naturalism I advocate in this essay. To adequately defend the claim that metaethics and moral psychology are in need of a thoroughly naturalistic turn, then, I will have to show that the Canberra Plan and similar approaches to moral inquiry are unlikely to make much progress toward understanding the nature of moral agency. To do this, I will look at their treatment of moral psychology and the problem of moral motivation in particular.

§6. An Outline of the Master Argument by Chapter

In Chapter 2 I begin with a careful examination of a recent dispute concerning the nature of moral motivation in metaethics, namely, the dispute surrounding the Humean theory of moral motivation. The Humeanism dispute, as I will refer to it, is represented by the third proposition in Smith’s inconsistent triad. It is therefore one of the foundational disputes in the moral problem and for the reasons just outlined a crucial part of the central organizing problem in
A substantial portion of the literature in metaethics is devoted to discussing the merits of two competing theories, namely, Humeanism and anti-Humeanism, both of which purport to accurately explain the relationship between moral judgment and motivation.

Humeans claim that motivation requires the presence of a moral judgment and a conative state such as a desire, while anti-Humeans deny this, generally because they think moral judgments are themselves sufficient to motivate. For these reasons the dispute can be used to illustrate how particular philosophical methods have been put to work in metaethics and moral psychology. Because many self-described naturalists have invoked empirical data from the neural and cognitive sciences to defend one of these two theories, the dispute also offers an opportunity to assess the promise of quasi-naturalistic approaches in metaethics.

Both Humeans and anti-Humeans trade exclusively in traditional, commonsense or “folk” mental state terms and concepts. This makes the application of empirical data somewhat contentious because the extent to which the sciences actually invoke such concepts directly remains a matter of controversy. It seems to me that vague treatment of this issue has caused some confusion in the literature.

Some moral philosophers who take a position in the dispute assume that the considerations relevant to settling a dispute couched in the language of commonsense will be commonsense explanations of behavior, and their methods make this assumption plain. Yet many self-described naturalists invoke empirical data, particularly from psychology and neuroscience, to vindicate one of the two commonsense theories. Despite professing an interest in the relevant empirical research, they evidently think that one of the two positions must ultimately be vindicated. This seems to carry the assumption that commonsense folk concepts
and terms *must* somehow appear in the relevant scientific research programs. But it is an assumption that few moral philosophers have bothered to defend first.

JPS are an important exception to this trend. They have defended the assumption explicitly, claiming that the mental state roles involved in folk psychological discourse must ultimately be realized in future neuroscience. Consequently, there seems to be at least two distinct ways to view the Humeanism dispute, and it is far from clear that these moral philosophers aren’t at times just talking past one another.

The primary contribution of Chapter 2 in general terms is therefore to draw what I take to be an underappreciated distinction between two distinct philosophical projects in the Humeanism literature. The first question facing philosophers engaged in the Humeanism dispute is whether they intend these theories of moral motivation to accurately describe folk psychological discourse and practice, i.e., to accurately describe what the folk say and do when they predict and explain each other’s behavior *regardless* of whether this commonsense discourse is ultimately empirically accurate or “respectable;” or, alternatively, whether they intend to develop an empirically respectable philosophical analysis of the plain facts about the relationship between moral judgment and motivation in real moral agents with real nervous systems.

In other words, I will argue that we can draw a distinction between engaging in the Humeanism dispute from the standpoint of the *folk perspective* and engaging in the dispute from the standpoint of the *philosophical perspective*. When we do the former, we avail ourselves of the resources that “ordinary folk” have at their disposal in predicting and explaining one another’s cognition and behavior. That is, we try to determine which folk theory of moral motivation, Humeanism or anti-Humeanism, is present when the folk think, talk, or discourse about what the folk do. This of course assumes that one or the other is actually being invoked,
which is to assume what philosophers call the “theory-theory”: the idea that folk psychology constitutes a legitimate theory. This project need not engage with empirical data except perhaps for data generated by the systematic observation of folk behavior.

When we engage in the latter project, by contrast, we avail ourselves of all the resources that come with contemporary philosophical inquiry, including empirical data in the case of (quasi-) naturalism, to determine which theory of moral motivation most accurately describes how moral agents really work, and not how people think, describe, or talk about how moral agents really work. These projects are generally conflated because it is typically just assumed that folk discourse can only be coherent to the extent that it tracks the actual empirical facts. But that assumption is far more controversial than it first seems.

This distinction between folk and philosophical projects in turn gives rise to a dilemma for philosophers who are in the business of developing theories of moral agency. The dilemma is roughly this: on the one hand, if disputes about moral agency, such as the Humeanism dispute, are genuinely about commonsense folk psychological discourse, we may find that we lack the resources needed to solve it. In this case the pertinent problem facing moral philosophers is to give legitimate reasons for thinking that the problem must admit of solution in favor of one theory or another. On the other hand, if the dispute concerns which philosophical theory of moral motivation is the most empirically accurate, the pertinent question is whether either theory really could be empirically vindicated.

I will argue that it is often very unclear which of these projects philosophers take themselves to be engaged in. For this reason it is difficult to determine which considerations—observational facts about commonsense psychology or empirical facts about nervous systems—are relevant to settling the dispute. In particular, I think this is true within the putatively cohesive
body of arguments advanced by the Canberra Planners, who sometimes talk as though they are concerned strictly with commonsense considerations about moral agency and sometimes talk as though careful philosophical work will reveal which of these theories of moral psychology is likely to be vindicated by future neuroscience.

Unsurprisingly to those familiar with the work of Jackson, Pettit, and Smith, this tension is explained by the fact that they take the conceptual analysis in the first step in the Canberra two-step to reveal a series of platitudes about the nature of beliefs and desires, platitudes employed implicitly (and perhaps sometimes also explicitly) by the folk, which must be preserved by future neuroscience. They argue that the functional, “folk roles” (this is a technical term to be discussed later) of belief and desire must appear in future neuroscience in order for folk psychological discourse, including discourse about moral agency, to be as successful as it is. This makes the project a thoroughly philosophical one, and not one concerned exclusively with a commonsense explanatory framework.

In more specific terms, then, the purpose of Chapter 2 is actually preemptive. In later chapters, Chapter 3 in particular, I offer a critical evaluation of the methods of the Canberra Plan on its own terms. The stock response to these criticisms is to point out that the Planners are merely interested in the commonsense project, but the claim is not supported by the methods typically employed. I want to prevent equivocation about the study of folk psychology by restricting this kind of vacillation between folk and philosophical projects. Though the first step of the Canberra method will amount to an analysis of platitudes, which is an analysis of commonsense concepts, it is nevertheless a philosophical analysis of human agency that avails itself of philosophical techniques. Once the myriad appeals to commonsense have been removed
and we get a better grip on what Canberra moral functionalism is not, we find ourselves in a better position to assess the approach given its own explicit methodological commitments.

In Chapter 3 I assess the state of some recent quasi-naturalistic approaches, including JPS’ version of the Canberra two-step, as philosophical research programs. In particular, I argue that these research projects, which are concerned with moral psychology and moral motivation especially, tend to violate their own methodological commitments. In the case of JPS, the putative conceptual analysis constitutive of the first step sometimes looks suspiciously like empirical psychology conducted a bit too carelessly from the armchair. I argue that the same is true of most quasi-naturalistic approaches, including some who do not subscribe to the Canberra Plan specifically. I will try to show that this is actually an unsurprising result. The kind of conceptual analysis that quasi-naturalists employ is likely not even capable of generating useful information about the cognitive processes of real moral agents. Consequently, armchair psychology is a natural place to look for the missing information.

In Chapter 4 I argue that one of the reasons that the kind of empirical observations used by JPS and other quasi-naturalists amount to shoddy psychology is a complete failure of consistent definitions of the key terms and concepts in the moral motivation literature. To use observational evidence to vindicate substantive theories of the nature of moral agency without first specifying the kinds of cognition and behavior definitive of the key terms in the theory, such as “judgment” and “motivation,” is in many cases just to beg the question in favor of one theory or the other. Properly, definitional disputes should come prior to disputes about the nature of moral agency. And at the very least philosophers must be careful to clearly define their terms prior to defending theories that invoke those terms.
Moreover, anecdotal evidence and hypothetical cases are inadequate substitutes for careful and systematic empirical research. Consequently, the theories developed by quasi-naturalists including JPS are not even amenable to empirical assessment as they stand. I conclude the chapter by offering some preliminary operational definitions (to use a controversial term) of key terms in the moral motivation dispute. The goal is to ready theories like Humeanism and anti-Humeanism for empirical assessment. If JPS are right that careful philosophical investigation reveals that the roles of belief and desire—which are the crucial elements of these theories of moral motivation—must ultimately appear in future neuroscience, then we will need a clearer idea of what to look for in the scientific literature.

In Chapter 5 I begin my argument against the empirical adequacy of philosophical theories of moral motivation, and so for the second horn of the dilemma, by reviewing and assessing some recent research programs in the neural sciences that seem to bear directly on moral judgment and decision. If JPS are right that beliefs and desires must appear in future neuroscience, then we can perhaps ask whether those roles already appear in the branches of neuroscience most relevant to the study of moral agency.

I therefore outline a new two-stage model of decision-making now emerging from "neuroeconomics," a relatively new research program which combines economic theory and cellular neuroscience in an effort to reveal the mechanisms of human economic behavior. I will argue that the model provides a useful context in which to understand a wide range of research programs in the neural and cognitive sciences. The decision-making model helps to unify these independent lines of research. I will argue that an extensive body of empirical literature is now converging on a fairly unified explanation of the mechanisms of moral decision via an extensive body of neuroscience literature on the contribution of emotion to decision-making. This novel
empirical synthesis begins the positive project, or constructive contribution, of this dissertation: a philosophically informed synthesis of existing lines of research with, I will argue, immediate implications for naturalistic projects in ethics.

Finally in Chapter 6 I complete the synthesis of empirical data and argue that the roles of belief and desire, even where they appear in the relevant neural sciences, are unlikely to be involved in the hypotheses that neuroscientists are already using to explain the mechanisms of human moral judgment and motivation. Though the case is, of course, not conclusive at this early stage, there is good reason to worry that the folk concepts are marginalized and explanatorily inert in the developing neuroscience of moral decision. Consequently, it is likely that both Humeanism and anti-Humeanism, in which these mental states play a central role, are both outmoded and irrelevant explanations of moral judgment and motivation. And this suggests the need for a more complete methodological naturalism in moral philosophy.
2. Moral Motivation and Philosophical Psychology

§1. Preliminaries

In the rest of this dissertation I assess the state of state of contemporary metaethics and ethical naturalism in particular by looking carefully at one of the central disputes concerning the nature of moral motivation. In particular, I focus on the methods that philosophers have been using to arbitrate the dispute surrounding the so-called Humean theory of moral motivation. Humeans claim that moral beliefs are insufficient to motivate since motivation requires the presence of an intrinsically motivating state such as a desire. Anti-Humeans deny this. The dispute is one of three core components of contemporary metaethics, which seeks to explain the nature of morality and moral agency.

I think that the dispute is deeply misguided. The arguments in the following five chapters generate a two-horned dilemma for those who hope to develop empirically respectable philosophical theories concerning the nature of moral motivation. As a dispute about the most accurate way to explain moral motivation from the standpoint of commonsense and/or a priori philosophical psychology it is insoluble; as an empirical dispute it is already obsolete.

The dilemma is intended to show that contemporary metaethics and ethical naturalism in particular have stalled because of widespread commitment to a backward, “quasi-naturalistic” methodology whereby they seek to solve traditional philosophical puzzles by drawing upon empirical observations and, in the case of naturalists, data emerging from the science of the mind and brain. We must reverse direction and modify even our most basic philosophical questions and theories about moral agency according to what we now know about the minds and brains of
real moral agents. I argue that the intractability of the Humeanism dispute is best explained by this broader methodological problem.

§2. The Humeanism Dispute

There is a widely discussed problem in metaethics concerning the causal relationship between two putative psychological states: the psychological state an agent is in when she forms a moral judgment and the psychological state she is in when she is motivated to act in accordance with that judgment. That the former plays some role in generating the latter is generally supposed to be a conclusion of common sense. More precisely how this works is generally supposed to be a conclusion of careful philosophical argument. The metaethics literature is now saturated with such arguments. All of them defend either of just two general positions.

Those in the first camp agree that a moral agent’s having a moral judgment—a belief about what it is right to do in a particular situation—is insufficient for her being motivated to act accordingly. They further agree about the reason for the insufficiency: they claim that in addition to her having a moral belief, she must also have a conative psychological state, that is, a separate psychological state that is intrinsically motivating in a way that beliefs are not. These additional conative states are generally taken to be desires. Proponents of this view are called “Humeans” because the view is supposed to have originated with David Hume, though this has been disputed (e.g., Millgram 1995). In any case the name has been retained.

Those in the second camp are united primarily by their disagreement with Humeans. They claim that an agent’s having a moral belief is sufficient to motivate, though they do not all agree on exactly how this works. Some think moral beliefs are sufficient to motivate directly,
and some only indirectly by, for example, producing a desire. Proponents of this view are called “anti-Humeans.” Anti-Humeanism is a broad term and sometimes encompasses narrower theories such as “rationalism” and “Kantianism,” which deny the Humean theory because they claim moral judgment and motivation derive from our rational capacities rather than our sentiments, though of course one could deny Humeanism without also being a Kantian or a rationalist. Here I follow Michael Smith (1994) in using the broader terminology both because I have an interest in his “moral problem” as the central organizing problem in metaethics and because the arguments that I offer in this essay are fundamentally about philosophical methodology so I intend them to apply mutatis mutandis to the narrower disputes, such as those about Kantianism, rationalism, so-called dual-process theories of moral agency, and the like, all of which are similarly situated in the tradition of a priori philosophical psychology.

Following Michael Smith we can express Humeanism more properly as the following proposition:

(H). An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume’s terms, distinct existences (1994: 12).

The term “Humeanism” is apt to cause some confusion because it implies one unified claim where there are actually three distinct claims: a claim about the relationship between three kinds of mental states, namely belief, desire, and motivation; an explicit commitment to the idea that beliefs and desires are distinct in kind; and the claim that beliefs and desires actually exist, which is, I think, entailed by the second. Call these the “insufficiency claim,” the “distinctness claim” and the “existence claim” respectively.

It is the insufficiency claim that is usually considered distinctive of Humeanism as a position on the problem of moral motivation. This is because it is a theory that is generally
invoked in discussions of folk psychology (hereafter “FP”), our rough and ready everyday understanding of cognition and behavior.\textsuperscript{17} FP just takes for granted that beliefs and desires exist as distinct kinds of psychological states. Humeanism, at least in this particular literature, is first and foremost a theory about the cause of motivation, but it includes an explicit commitment to the distinctness of beliefs and desires, which in turn entails the existence of those states.\textsuperscript{18}

The distinctive feature of anti-Humeanism, by contrast, is its sufficiency claim, i.e. that moral beliefs are themselves sufficient to motivate. Importantly, however, some subscribe to the sufficiency claim, and hence anti-Humeanism, because they think it false that beliefs and desires are necessarily \textit{distinct} psychological states. For example, some think that beliefs and desires, at least when it comes to an agent’s forming moral judgments, can be two parts of one mental state, or alternatively two mental states that always co-occur (e.g., Little 1997, McDowell 1978). So there are generally three different reasons to adopt anti-Humeanism. An anti-Humean may claim that moral judgments are themselves sufficient to motivate because: (i) moral judgments are expressions of desires or similar conative states (“expressivism”); (ii) moral judgments produce or co-occur with desires or similar conative states (which may amount to “judgment internalism”); (iii) moral judgments are unified, individual states that are simultaneously belief-

\textsuperscript{17} FP will sometimes also stand for “folk psychological.”
\textsuperscript{18} It is plausible, however, to read Smith as claiming the reverse, namely that the distinctness of beliefs and desires is what explains the theory of motivation. He seems to conceive of the dispute between Humeans and anti-Humeans as a dispute about whether we are just believers and desirers or believers, desirers, and “besirers” as well (cf. 1994: 120). That is, as a dispute about whether beliefs and desire can always be pulled apart, at least modally. I’ll have much more to say about this later. Here I just want to point out that we can distinguish between the claim about motivation and the claim about the distinctness of beliefs and desires. Much of the discussion in the literature seems to take the theory of \textit{motivation} as primary to Humeanism and the distinction as entailed. Whether Hume, the purported source of these ideas, developed it in the opposite direction is a historical question without any real bearing on my thesis, and one that takes for granted that developed the theory as it is used today, which is doubtful.
like and desire-like (i.e., they are what some philosophers call “besires”). There are two points to notice about the dispute.

§3. Belief/Desire Realism

The first point has to do with the status of beliefs and desires. It emerges when we consider which of the Humeans’ specific claims the anti-Humeans have rejected. Anti-Humeans who claim (i), i.e., expressivists, reject the insufficiency claim since desires are thought to be like conative states. If moral judgments just are desires then there will be no mystery about how such judgments motivate since judgments qua conative states are motivating. They certainly accept the existence claim since they take judgments to be desires, though what they say about the distinctness claim will vary. In general most probably accept it: if judgments are desires, they are not beliefs.

Anti-Humeans who claim (ii), i.e. that moral beliefs produce or co-occur with desires, also object to the insufficiency claim, though they do so because they deny the distinctness claim. They claim that if we have a moral belief a desire will follow, either because the former produces the latter or because one somehow simply cannot occur without the other. Moral beliefs are sufficient for motivation because they are closely linked to desires. But they clearly accept the existence claim: if these states co-occur, or if one produces the other, then they must exist.

Anti-Humeans who claim (iii) present a slightly more difficult case because the concept of a single unified state that is both belief-like and desire-like, i.e. the concept of a “besire,” is less familiar or commonsensical. The use of the phrases “belief-like” and “desire-like” are

\[19\] The term “besire” is due to J.E.J. Altham (1986).
informative, though, because they work by referencing the definitions of belief and desire in the philosophical literature. These in turn are generally cashed out in terms of their “directions of fit,” a concept due to Anscombe (1957). Beliefs are mental states that fit the world, i.e. they have a mind-to-world direction of fit, while desires are states with which the world must be made to fit, i.e. they have a world-to-mind direction of fit. As Smith says, our desires tell us how we want the world to be while our beliefs tell us how the world really is so that we can go about changing it (1994: 9).

The metaphor is supposed to capture our commonsense understanding of these FP concepts, and perhaps Hume’s too, and so it is one that has come to occupy an important role in the moral motivation literature. Some versions of anti-Humeanism seem to depend on whether it is possible for a single mental state to have two opposing directions of fit, which has given rise to dispute (e.g., McDowell 1978; Little 1997; Smith 1994). And Smith has offered a technical distinction on the basis of Anscombe’s metaphor. He proposes to distinguish between beliefs and desires on the one hand and belief-like states and desire-like states on the other by claiming that the former are capable of having only one direction of fit while the latter are capable of having two (1994: 117). Whatever else we might make of these claims I take them to illustrate the extent to which directions of fit have become philosophically entrenched.

At least to the extent that the metaphor is entrenched, Anti-Humeans who claim (iii) seem to take seriously the idea that mental states can be distinguished in some such terms. That philosophers have already done work to differentiate between beliefs and desires is presumably what makes the reference to those states more useful than simply choosing names that don’t refer to already familiar mental state concepts. To say that a state is like a belief or like a desire is to
say that it has some of the properties we usually ascribe to a belief or to a desire. And this is just to point toward the stock philosophical definitions. Here, for example, is Smith’s:

For the difference between beliefs and desires in terms of direction of fit can be seen to amount to a difference in the functional roles of belief and desire. Very roughly, and simplifying somewhat, it amounts, \textit{inter alia}, to a difference in the counterfactual dependence of a belief that \( p \) and a desire that \( p \) on a perception with the content that not \( p \): a belief that \( p \) tends to go out of existence in the presence of a perception with the content that not \( p \), whereas a desire that \( p \) tends to endure, disposing the subject in that state to bring it about that \( p \) (1994: 115).

So anti-Humeans who claim (iii) clearly deny the insufficiency claim and the distinctness claim, but they seem to accept the existence of beliefs and desires in addition to besires.

Anti-Humeans of all stripes share with Humeans of all stripes a commitment to the view I will call belief/desire realism (or FP realism).\textsuperscript{20} In this I follow (among others) Horgan and Graham (1991) who explain FP realism as the view that “our everyday FP descriptions of people are by and large true, and thus that humans generally do undergo the FP events, beliefs, desires, and so forth that we normally attribute to them” (107). Both Humeans and anti-Humeans claim that moral agents have psychological states adequately characterized by the folk concepts of belief and desire, and perhaps also desire, and that we can at least distinguish between these states in theory, even if not in practice.

This result should be uncontroversial. The distinctive feature of the Humean theory is the sufficiency claim rather than the explicit distinctness claim because the dispute over Humeanism is generally situated within the context of commonsense FP, or as Smith says, the “standard picture” of human psychology (1994: 11). No one who rejects FP will have any stock in the outcome of a dispute about the causal efficacy of the mental state types it postulates. This is of

\textsuperscript{20}I will use these terms interchangeably in this dissertation and I take no position on the question whether these should be distinguished.
course why eliminativism—the view that beliefs, desires, and the like do not really exist—is not a form of anti-Humeanism.

One obvious objection here is that there are many different accounts of FP, some of which carry no metaphysical commitments about its postulates. Instrumentalists (e.g., Dennett 1978, 1981), for example, take these putative mental states to be something like useful fictions—useful for the purposes of prediction and explanation of behavior, but not (necessarily) metaphysically real entities in the head. Dennett famously claims that attributions of belief and desire are really just “idealized fictions” in a calculus that explains action. They are abstracta that do not correspond to any physical or functional state in an actual organism. Perhaps some who engage in the Humeanism literature could take this line and thereby avoid any commitment to belief/desire realism.

I doubt this because Humeanism and anti-Humeanism are theories about the causal efficacy of putative mental states, i.e., theories about the effects that moral beliefs and desires have on an agent’s motivation. I take it as obvious that a dispute about the causal effect that some mental entity has on another mental entity or psychophysical process presupposes the existence of those mental entities. And in any case the distinctness claim—“where belief and desire are, in Hume’s terms, distinct existences”—is an explicit existential commitment, which anti-Humeans must accept lest the position becomes indistinguishable from eliminativism.

Both, then, are realists regardless of their broader positions on the status of FP, such as whether they subscribe to “East Coast Realism” or “Southern Fundamentalism,” to borrow from the lexicon of Horgan and Graham (1991). Settling the dispute will require explicating the

---

21 Roughly, both East Coast Realists and Southern Fundamentalists are realists about folk psychology, but the Easterners are apologists in the sense that they accept the existence of falsification conditions and then attempt to argue that those conditions have not been met while
roles of these commonsense psychological states in generating moral motivation, which assumes that these states are real enough to figure into generalizations about moral agency, and real enough that we can distinguish between them in principle. This seems to be the point when, for example, Smith says the philosopher’s task is to make sense of “ordinary moral practice as it is engaged by ordinary folk” (1994: 5). And Nichols says that “it is clearly the case that lay people know a number of platitudes about morality” and that “the project of charting those platitudes is both practicable and interesting” (2002: 286).

Stephen Stich makes a similar point about instrumentalism, though his project is concerned with philosophy of mind rather than metaethics or moral psychology. He rightly observes that, “we often talk about beliefs and desires as though they had causal properties” (1983: 243). Stich goes on to provide toy examples. If you give the police a bad tip about your friend’s whereabouts, you will later claim when pressed that you gave the bad tip because you genuinely believed your friend was just where you had said she was.

Here we can trade the toy example for some real examples from contemporary metaethics. Philosophers who engage in the dispute about the Humean theory of moral motivation are in dispute about whether moral beliefs are sufficient for motivation. Spelled out more fully, they are in dispute about whether moral beliefs are sufficient causes of moral motivation. The entire dispute, it must be emphasized, is about the causal efficacy of putative mental states. Those who intend to participate in a non-causal version of the dispute should be explicit on that point.

the Southern Fundamentalists argue instead that evidence for the satisfaction of those falsification conditions could really only serve to show that the conditions themselves are the wrong conditions to demand.
Some participants have been explicit about their philosophical commitments on the metaphysics of mental states (though generally in places other than the Humeanism literature) and the causal relevance of those states. But the result is sometimes more confusing than clarifying. In particular, Frank Jackson, Philip Pettit, and Michael Smith (2004; hereafter “JPS” when I refer to their collective work) seem to be the core of a philosophical program that among other things defends the viability of intentional moral psychology. Pettit and Smith (2004) open their argument for a particular account of desire with this: “the intentional conception of human beings is endorsed by philosophers on most sides, if not quite on all, and we start from the assumption that it is sound” (270). They pin to this a footnote in which they point out that eliminativists like Churchland and Stich exhaust the minority opposition, and conclude: “However, that the intentional conception is sound need only mean, for our purposes here, that it is the sort of useful fiction which D.C. Dennett takes it to be” (270n1).

It seems we are already in disagreement. For I have just said that the only way to make sense of a theory concerning the causal efficacy of a mental state is to take for granted that the state in question exists. It seems reasonable to suppose that only things that physically exist can actually cause anything physical. For example, there is probably not much use in formulating a theory of internal combustion engines if we think that gasoline may turn out to be a useful fiction. Yet despite their interest in defending common sense, JPS seem to think this observation is philosophically naïve. Jackson and Pettit (1990a-c) have tried to show that functional properties are causally relevant.

I will return to this issue shortly. Though this preliminary observation about the Humeanism dispute is evidently more controversial than I thought, it was Smith and not me who formulated Humeanism in such a way as to make it explicit that beliefs and desires are distinct
existences. In fact Smith’s account of the belief/desire distinction is steeped in talk of existing and enduring. Smith could have said instead that beliefs and desires are distinct useful fictions, or maybe just distinct concepts. But these formulations sound considerably less Humean. The footnote notwithstanding, though, JPS are clear that they think the “intentional conception of human beings” is sound and that most would agree.

§4. Humeanism and the Moral Problem

The second point I want to make about the Humeanism dispute concerns its broader role in metaethics. It is closely connected to metaethical disputes about cognitivism and judgment internalism. To see why this is so, consider Smith’s (1994) “moral problem.”

Smith has observed that the deep disagreement characteristic of contemporary metaethics is best explained by the tension between three widely accepted ideas about morality, two of which we intuit and one which we supposedly inherit from Hume. First, we take moral judgments to have objective purport; moral disagreement arises from the natural inclination to express our moral judgments as though they were objective facts. When I say that torturing calves for the production of tender veal is wrong, I don’t mean that I happen to dislike it and that you are nevertheless entitled to your own opinion about the matter. I mean that there is something about the way the world is such that the practice is wrong whenever and wherever it occurs, and that you would agree with me if you were a decent person. So moral judgments seem as though they must be representational mental states, that is, beliefs.

Second, we take those judgments to have a special connection with motivation. They are fundamentally action guiding. What makes a moral judgment different from an everyday, ordinary judgment about tables or chairs is that the former seem tied up with the inclination or
desire to *do something* about them, such as, to take the same example, boycotting veal. So moral judgments are fundamentally practical. But if moral judgments are to be practical for each of us, they must be expressions of our own subjective desires – expressions of what we *want* to do.

But it seems that moral judgments cannot be both expressions of belief and expressions of desire. For the third idea that we all share is nothing could be both the expression of an objective belief and a subjective desire. This is the basic picture of belief/desire psychology, which we supposedly inherited from Hume, according to which beliefs and desires are distinct existences and according to which motivation requires both states (Hume 2000, 2.3.3: 67; cf. Smith 1994: 7). The tension is thus between the objectivity and practicality of moral judgments that arises within the “Humean” theory. This is the “moral problem.” And the last half-century of work on that task has resulted in little progress among moral philosophers (Darwall et al. 1997; Smith 1994).

This problem can be thought of as the central organizing framework of contemporary metaethics because it helps to explain the widespread disagreement among philosophers about metaethical issues.22 This framework—the inconsistency in our metaethical intuitions—can be neatly illustrated with the following three apparently inconsistent propositions:

1. Moral judgments of the form ‘It is right that I ɸ’ express a subject’s beliefs about an objective matter of fact.

2. If someone judges that it is right that she ɸ then, *ceteris paribus*, she is motivated to ɸ.23

---

22 This is Smith’s claim, but I agree with him. As I argue in the previous chapter the many different metaethical positions appear to grow out of the need to resolve this tension. It is a claim about which issues were a part of the historical development of metaethics rather than a claim about the influence of Smith’s particular formulation.

23 There is a dispute between Smith and David Brink (1997) about whether this proposition adequately captures the commitments of the noncognitivist. According to Brink, noncognitivists claim that moral judgments entail motivation. In other words, Brink claims the *ceteris paribus* clause dilutes the proposition too much to capture the idea behind noncognitivism. This sets up
3. An agent is motivated to act in a certain way just in case she has an appropriate desire and a means-end belief, where belief and desire are, in Hume’s terms, distinct existences.

Though none of them are entirely beyond controversy, Smith claims that proposition (1) expresses cognitivism, (2) judgment internalism, and (3) Humeanism or the Humean theory of motivation, with its associated claims, that is the subject of this chapter. The apparent inconsistency arises because (1), (2), and (3) together entail a necessary connection between moral beliefs and desires, which are distinct existences, while (3) denies that very connection. Moral judgments express beliefs (“cognitivism”), which according to internalism are necessarily connected to motivation, and Humeanism tells us that it is desires that provide that connection. From this we get the necessary connection between beliefs and desires, but it is this connection that Humeanism explicitly denies in claiming that beliefs and desires are distinct existences.

Put in this more technical way, the “moral problem” is the problem of determining which intuitively appealing theory of moral agency to jettison from the triad: cognitivism, internalism, or the Humean theory of motivation. So it should now be clear why Humeans seem compelled to accept cognitivism: to accept (3) without (1) is to require the presence of beliefs for the production of motivation without leaving any real work for them to do. If motivation requires a belief and a desire, which are distinct, and if it is the desire that is both motivating and also the expression of the moral judgment, then it is a mystery what the belief has to do with anything. In other words, the Humean theory of moral motivation is only really a meaningful theory of moral motivation insofar as it is committed to the idea that moral judgments purport to state facts about

---

his criticism of Smith’s solution to the moral problem, which involves making the three propositions compatible. Brink’s point is that Smith’s solution works only because it uses a weak formulation of internalism. Smith (1997) offers a few reasons to prefer his own phrasing. I don’t intend to enter into the dispute here. I adopt Smith’s phrasing rather than Brink’s only because it is the original formulation.
the world, i.e., express beliefs. For, again, if those judgments express desires then any further attempt to explain their link with motivation will be redundant.

The crux of the issue is the status of beliefs and desires. The Humean theory with its commitment to belief/desire realism is the centerpiece of Smith’s inconsistent triad. It provides the psychological framework from which we must accommodate both the subjectivity and objectivity of moral discourse. Without any commitment to a psychological theory according to which beliefs and desires are distinct existences there isn’t any tension (cf. Smith 1994: 7).

To see this point, consider what’s at issue in each of the theories represented by these propositions. In one sense, cognitivism and internalism—at least as expressed in Smith’s framework—are substantive issues only because we assume the truth of the distinctness claim expressed in the Humean theory, while Humeanism with its belief/desire realism is substantive in its own right.

Cognitivism according to Smith is the view that moral judgments express beliefs about objective matters of fact. There are two things to notice about this claim. The first is the “objective matters of fact” clause and the second the “express beliefs” clause. Most of us probably don’t think much about why the former is even included, and would probably read the claim the same way even if it weren’t. This is because we just associate beliefs with facts automatically. Beliefs are about facts; they are fact stating mental states. Similarly, we just associate desires with values. And we see these pairings as diametrically opposed. So long as we make this assumption the issue is clear. Either moral judgments are beliefs which purport to
express facts or they are desires which express something like emotions or values. If we disrupt these associations we also disrupt the dispute.

Suppose that empirical investigation reveals that the mental states we currently call belief and desire are actually two inextricable components of one mental state or process. This one state is really an emotional state or feeling that serves to represent the way the world is, or to help the organism navigate its environment, which includes both its internal and external states (or something like this). For lack of a better term, we might call this newly discovered mental state a ‘besire.’

In this case, (1) would read differently. It would say that moral judgments express a subject’s besires about objective matters of fact. The claim is true by definition, and it is no longer clear what the opposing view might be.

It cannot be the claim that moral judgments are really expressions of besires about objective matters of fact since that is exactly the same claim. The alternative possibility seems to be the idea that moral judgments are expressions of a subject’s besires, which are not about objective matters of fact. Perhaps they are expressions of besires about values, which are subjective. But what could this mean, given that the mental state in question includes a feeling that serves to represent the way the world really is? The dispute now seems confused because

---

24 Of course you can have beliefs about desires, but these are still beliefs because they are states characterized by assent to propositions about desires. The point is that it would be incoherent in our widely accepted framework to claim that beliefs are characterized by assent to desires themselves.

25 The term is due to Altham (1986) and is already a major part of the literature in metaethics. I am pretending otherwise and using it very differently here just for the sake of this example. In the Humeanism literature besires are taken to be mental states having both “directions of fit”: they are simultaneously states that fit the world and states with which the world must fit. You might say, if you insist upon using the metaphor, that besires as used here have only one direction of fit: they fit the world, though they do so bimodally (both cognitively and emotionally).
the opposing view is at odds with the stipulated definition of a besire, which is an emotional state that serves to represent the way the world actually is.

So the debate about cognitivism might actually depend upon the existence of beliefs, which we associate with the expression of facts and not values. Because we are accustomed to working within the folk psychological (FP) framework according to which beliefs and desires are distinct existences, we tend to read this distinction right into (1). This is one sense in which I think Humeanism dominates the metaethical framework. And this may help to explain why the debate about cognitivism, at least within the context of the moral problem, appears so intractable. If the dispute surrounding Humeanism rests on false assumptions then there will be serious implications for the dispute about cognitivism.

A similar point holds for the internalism debate, which is proposition (2) in the inconsistent triad. It is a dispute about the nature of the connection between moral judgment and motivation. According to judgment internalism, moral judgments motivate necessarily.

---

26 This might raise the following worry about Smith’s particular formulation of the cognitivist thesis. The claim in this paragraph might not hold if we adopt the non-mentalistic formulation of the cognitivist thesis, which does not refer to beliefs or mental states directly. Cognitivism on this formulation is the view that ethical sentences express propositions and are therefore truth-apt. But if this is so, then the inconsistency in Smith’s framework might just be an artifact of the peculiar, mentalistic way of formulating the view.

Here is one possible reply on Smith’s behalf. Cognitivism is called such because it claims that ethical sentences are cognitive. That is, ethical sentences are meaningful representations of mental content. Formulating the view without mentalistic language, then, does nothing to change the fact that it is foundationally a claim about mental content and not about sentences. One might still object that the important part of (1) is the “capable of being true or false” clause, but then this is really just to shift our concern to whether there are actually truth-makers for ethical propositions, which is the debate about moral realism, not cognitivism. In other words, the cognitivism debate is really a debate about the relation our human moral discourse stands in to ethical properties rather than about the metaphysics of those properties themselves, which is the question of moral realism.
Judgment externalism is the denial of this view. Here again the dispute hinges on our assumptions about beliefs. This is to be expected given that the debate about internalism overlaps substantially with that of Humeanism. Humeanism concerns whether moral judgments as beliefs are sufficient to motivate. Internalism concerns whether moral judgments motivate necessarily. But both disputes concern the causal efficacy of moral judgment. And in both cases, one’s commitments are determined by one’s prior account of the nature of moral judgments, where this involves choosing between the usual options.

If you think a moral judgment is a belief in the standard (Humean) sense of the term (i.e., a mental state that represents the way the world is) then you will probably have no reason to claim that such beliefs necessarily motivate. You probably think that moral judgments are some kind of coldly rational, cognitive, world-representing state quite distinct from the kind of states that are rich in emotional content, feeling and sensation, and it is the latter that intuitively drives you to want to do something. As we saw in the first chapter, this observation is what gives rise to the fundamental dilemma of metaethics in the first place. And if you think just the opposite, that moral judgments are expressions of feelings, the kind of state opposed to the coldly rational representation of the world, you will have a hard time claiming that they are not sufficient to motivate for the same reason. The traditional dispute about the nature of moral motivation is now underway.

But suppose that it has been determined that some of the states we once thought beliefs or desires are really besires, which we have defined as an emotional state or feeling used by an organism to represent the way the world is (and perhaps to help it navigate its environment, etc.). The claim is then that moral judgments are besires. The question becomes: what obvious facts
about the causes of motivation can we derive from reflecting on emotion-based representation of
the world?

With the usual ideas about beliefs and desires as mental states with distinct directions of
fit disrupted—which as I said previously at least appears to be an assumption shared by
Humeans and anti-Humeans alike—we are deprived of the litmus test by which we determine
our commitments. Desires on the usual picture are assumed to be conative precisely because
they are not about what the world is but rather what we want it to be. Without that background
assumption we have no way to answer the question. In order to settle the matter, we will need to
know much more about the nature of these besires. Just what kind of feeling is involved in
them? How are these feelings related to motivational states?

Intuition has relatively little to offer on these questions. But as things stand now moral
philosophers argue in great detail about internalism and the cognitive requirements for moral
motivation based on assumptions about mental states that they attribute to (strangely enough)
either common sense or David Hume. Smith says of the former: “what else do we have to go by

Once we accept that beliefs represent the world as it is and desires represent the way we
wish the world to be, we can sit back and reflect on what interesting claims about moral agency
follow from this. Without the assumption we are forced to ask more fundamental questions
about human mental states and agency. The philosophical dispute over proposition (2)
presupposes a controversial psychological framework.

The third proposition in the inconsistent triad differs from the previous two in that it is
substantive and interesting in its own right. It is a thesis about the nature of human moral
agency. It is the assertion of a psychological framework, not a question about the implications
for moral agency of accepting a psychological framework. It tells us that beliefs are mental states wholly distinct from desires. And when we consider the concept of a desire (in my sense of the term) what we get isn’t confusion but genuine controversy. Humeanism tells us that beliefs and desires are distinct existences. The possibility of desires suggests otherwise. One of these claims must be false and, as Hume himself recognized when he deferred to the anatomist on the origin and nature of mental states, armchair reflection won’t tell us which it is. All of this indicates just how entrenched belief/desire psychology has become in our metaethical framework. That psychological framework has turned metaethics into a dispute over the preferred set of pre-packaged commitments.

One way to avoid these pre-packaged commitments is to remove the tension in the moral problem rather than accommodate it. The way to do this is to recognize that the disputes expressed in its propositions are better dissolved than resolved. I will try to show that this is true of the Humeanism dispute in particular.

§5. Toward a Dilemma for Metaethics

I think that the dispute in contemporary metaethics surrounding the Humean theory as a theory of moral motivation is insoluble. My argument for this claim takes the form a dilemma, the two horns of which I defend over the rest of this dissertation.

27 Like Smith I want to reject all three of the “standard solutions” to the moral problem. That is, I follow Smith in rejecting the idea that the right approach is to give up one of the three apparently inconsistent propositions. I differ from Smith in that I want to explain how suspect philosophical methods have given rise to the moral problem in the first place. I’m not interested in removing the tension. Smith’s goal is to solve the moral problem by offering a revised understanding of Humeanism, one in which we preserve the distinction between the roles of belief and desire in moral motivation but reject the Humean account of the rationality of desire (1994: 129).
We can understand the dispute generally in only one of two ways, neither of which is likely to vindicate either position. If the dispute is a philosophical one about the relationship between mental states in the framework of a priori intentional psychology then it is a dispute closed to empirical solution and therefore, I will argue, to any solution at all. I will try to show that the methods employed by traditional analytic moral philosophers to settle the dispute tacitly concede this point. In this case, these philosophers are wrong to appeal to empirical observations about the behavior and cognition of real agents to settle the matter, which, as I shall show, they regularly do.

If, on the other horn, it is a dispute that purports to track our current science of cognition and behavior then it is already obsolete. There is a burgeoning body of literature in neuroscience already sufficient to show that the neurological mechanisms of judgment, decision-making, and motivation are far too complex to be adequately represented by either of these crude philosophical positions. The science supports neither Humeanism nor anti-Humeanism. The argument for this dilemma will take us into two different directions. An extensive review of recent empirical work in later chapters will be crucial to establishing the second horn of the dilemma. The argument for the first horn in the rest of this chapter and the next will take us into the philosophy of mind and philosophical psychology.

The dilemma is intended as one part of a broader argument for a genuine methodological ethical naturalism. This is roughly the view that ethical inquiry and scientific inquiry are continuous (Darwall et al. 1997). My primary goal in this dissertation is to show that the dilemma arises from the backwards methodology employed by contemporary moral philosophers, and by self-described naturalists especially.
Metaethics has much to gain from a genuine methodological naturalism because some of its disputes are insoluble as they stand. Self-described naturalists are wrong when they force empirical data from the study of the mind and brain to square with traditional philosophical puzzles about human moral agency, yet almost all versions of naturalism engage in that practice. In particular, the practice is problematic because it leads us to confront the physiological constraints placed on theories of morality by our neurological constitution only after we have established the philosophical problem. Consequently, many of these traditional metaethical problems are outmoded long before we even turn to the data. When we commit ourselves in advance to a particular problem and look for data to solve it we have already given up the crucial commitment of the naturalistic project. We should bring our philosophical questions into alignment with our empirical study of real moral agents rather than the other way around. Owen Flanagan (1991) once put this straightforward point in terms of the “principle of minimal psychological realism” (PMPR). But even those who heeded it in the subsequent two decades that saw the rising popularity of philosophical naturalism have failed to appreciate or implement the methodological shift that it requires.

I am in no way claiming in this essay that philosophers are wrong to argue about philosophical definitions of folk concepts, such as moral judgment, motivation, belief, and desire. Rather, if these are legitimately discussions about folk concepts, then disputes about their definitions amount to disputes about which proposed definitions best reflect the concepts that the folk implicitly or explicitly employ in everyday discourse. And this is not the same as a dispute about how best to define these concepts for the purposes of forming explanatory generalizations about what the folk (or moral agents) do when they actually undergo these putative mental states. How philosophers and scientists explain what the folk do and how philosophers explain how the
folk explain what the folk do are two different issues, which I think are frequently conflated (whether intentionally or unintentionally). Concerning the latter, it is an open question whether the folk really even explain what the folk do in terms useful to philosophical moral psychology and to the dispute about the Humean theory of moral motivation in particular. If not, then we will have no choice but to interpret the dispute as one about the theory-theory in moral contexts and many philosophers will be forced to revise some of their claims about commonsense and the folk. The rest of this essay follows each of the horns in Figure 2.1 from left to right.

![Diagram](image)

**Figure 2.1.** A Proposed Dilemma for the Humeanism Dispute in Moral Psychology

One of the keys to my argument is the idea that we can adequately distinguish between the two projects I have just described. The distinction between folk and philosophical projects is
marked in the diagram by the division of the first horn into two parts. The crucial claim will be
that because so many moral philosophers have been equivocating on the nature of their project,
or vacillating between these two projects freely, serious methodological problems have gone
unnoticed. By holding each of the projects constant before offering an assessment we can better
understand why a turn to methodological naturalism is long overdue. I turn now to the argument
for the first half of the first horn.

§6. Of Commonsense Theories of Moral Motivation

I said before that our commonsense understanding of mental states is called “folk
psychology” or “FP.” FP is sometimes thought to constitute a theory of mind. Because to hold
this view is to assent to a theory about FP, namely the theory that FP is in fact a theory (as
opposed to, say, a tacit understanding), philosophers call it the “theory-theory.” Stich &
Ravenscroft (1994) further distinguish between two different senses of FP: an internalist and an
externalist sense. Where FP is understood as a theory implicit in our everyday talk about mental
states it is external FP. Where it is understood to be a theory represented in the mind-brain that
serves as the foundation of our capacity to predict and explain behavior it is internal FP. Internal FP is supposed to be a kind of knowledge representation (Ravenscroft 2010). This
distinction can in turn be distributed over the distinction between FP and the theory-theory, so
that talk about FP falls under one of four categories: FP internal, FP external, theory-theory
internal, and theory-theory external. Ravenscroft has rightly observed that the internal/external
distinction has not caught on and so it is frequently unclear which meaning is intended in the
literature. For the same reasons I will avoid the proposed convention.
The idea that the dispute about the Humean theory of moral motivation falls under the province of philosophical moral psychology might be understood in two different ways depending upon the nature of our interests. On the one hand it might mean that the dispute is properly a dispute about how the folk actually understand the mental states involved in moral agency. On the other hand the dispute might be understood as a special case, concerning moral cognition and action, in the philosophy of mind. The distinction is illustrated in Figure 2.2 below. In fact there are four potential projects here, but at present we are concerned with just these two.

The first amounts to the idea that the Humeanism dispute is one about how to understand folk practice or discourse in moral contexts. In this case the philosophical project involves accurately characterizing or describing what the folk say about moral motivation on their own terms. The goal in such a project is to make sense of folk practice as folk: “What are we, the folk, doing when we make moral judgments?” This is P1 in the figure below.

On the second view the dispute is about how to understand the theory-theory in moral contexts – that is, as a special, moral case of the theory-theory dispute. We as philosophers are out to make sense of folk practice in philosophical terms: “What are they, those folk, doing when they talk about moral judgment, and how might we best systematize their discourse or practice?” This is P2 in the figure below.
**Figure 2.2.** Two Perspectives on Folk Discourse.

The figure raises at least the following four issues. First is whether these distinctions correspond exactly to the distinctions between so-called “analytic functionalism” (or commonsense functionalism) and “psychofunctionalism” (or a posteriori functionalism), which are already a part of the literature in the philosophy of mind. It is possible that what I am calling P1 may just be analytic functionalism and what I am calling P2 may just be a posteriori functionalism. I want to avoid that terminology in part because I think it glosses over the distinction between the putative source of the analysis, i.e., the difference in folk and philosophical perspectives. Even if I am wrong about this, I prefer to avoid the connotations that come with terminology already entrenched in the philosophy of mind.

The second issue is whether F2 is a possible project. This will depend on whether we think that FP as a body of knowledge constitutes a legitimate theory. Those who say no
essentially deny that F2 is even possible. I will eventually defend the view that FP is not a
theory.

The third is whether F2 and P1 are actually the same. I take it that they are roughly the
same project, though they differ in who carries out the analysis. In F2 it is the folk and in P1 it is
the philosopher. But given that both are committed to using exactly the same tools and methods,
the projects are for all purposes identical. The very idea of P1 might be strange, but it is
nonetheless a possible project so long as we accept the existence of a theory-theory.

A final question concerns the distance between P1 and P2. The projects are no doubt
closely related. The crucial difference between them as I see it concerns the resources available
for the analysis: the folk have common sense and perhaps “lore;” the philosophers have a more
systematic body of literature and technical devices and some training on how to employ them.
There is nothing strange or problematic about the idea that these projects must be closely related.
Philosophers are themselves just folk. Conversely, the folk can philosophize too.

Despite the similarities, P1 and P2 are distinct projects for precisely the reason that F2
and P1 are nearly the same. The difference in resources available for carrying out the project
must not be overlooked: from the folk perspective we should have only commonsense and
intuition at our disposal while in the philosophical perspective we should have all the tools and
background philosophical knowledge on mind and mental states to sort out the precise nature of
folk discourse.

I hope this claim is uncontroversial. If we want to accurately make sense of folk terms
and concepts as the folk use them then we must confine ourselves to the resources that they have
at their disposal. Compare folk physics: there is no use in purporting to describe naïve
interpretations of physical space if we are going to attribute assumptions rooted in complex
concepts in the physical sciences, like special relativity, *unless* we can first show how the folk in fact put those concepts to work, either implicitly or explicitly. But the lesson we are generally supposed to take from appeals to folk physics is that they don’t.

Perhaps one of the reasons that this point is ignored is that it seems condescending to drive a wedge between the folk and philosophers. Philosophers like to deny that they have to look down from their ivory tower. We can be sympathetic to that worry and nevertheless recognize a legitimate distinction. The profession depends on it. There is a difference, though no doubt it is one of degree and not kind. The objection from humility strikes me as disingenuous.

As a first step, then, we will need to determine whether all or most of the philosophers who engage in the Humeanism dispute take themselves to be engaged in the same sort of philosophical project. In which of these four projects do they claim to be interested? That is, are all Humeans and anti-Humeans in agreement about the nature of our philosophical interest in FP? If not, it is unsurprising that there is so much disagreement since in many cases moral philosophers will simply be talking past one another. If so, we’ll need to know which of these projects is really at issue if we hope to settle the dispute.

I think that moral philosophers involved in this dispute are not in agreement on the nature of their collective project because the distinction, simple though it is, is so frequently ignored. The terminology in the literature is inconsistent on the matter. It is surprisingly unclear how much philosophers intend to preserve from folk discourse.

I will start with Michael Smith (1994), since he opens his influential book with some remarks about the relationship between the folk and the philosopher:

Philosophers are ordinary folk… Of course, philosophers tend to give more technical and systematic answers to normative ethical questions than ordinary folk… The important
point, however, is that philosophers’ theories do not generate answers that are different in kind to the answers ordinary folk give to moral questions. They are merely more technical and more systematic. Despite their interest in normative ethics, however, philosophers have not tended to think that these sorts of questions are of the first importance in moral philosophy... Rather they have thought that we should do normative ethics only after we have given satisfactory answers to certain questions in meta-ethics (2; italics are Smith’s).

It is not exactly clear what Smith takes the implications of this to be for his own philosophical project in later chapters. On the one hand it seems that he takes the relevant difference between the folk and the philosopher to be little more than a difference in their ethical interests. The folk are interested in normative questions about what we ought, morally, to do. Philosophers are much more interested in what our claims about these matters really mean. If so, folk wisdom might be similarly capable of uncovering the truth about metaethical questions, like those about the nature of moral motivation. Alternatively, though, he might mean that the difference in interests between the folk and the philosopher is a crucial one. What the folk lack is the (mere?) technical expertise necessary to understand that and how broader philosophical questions ultimately suffuse their normative disputes with meaning.

There are reasons to think that he means something more like the latter. He follows this introduction with a protracted discussion of what he calls the “standard picture” of human psychology. He opens with the claim that metaethical questions engender so much disagreement because of two distinct features of morality “that are manifest in ordinary moral practice as it is engaged by ordinary folk” (5). And then the key claim: “The philosopher’s task is to make sense of a practice having these features.”

So according to Smith the Humeanism dispute, like all metaethical disputes, is an attempt to make sense of ordinary practice, i.e., ordinary moral discourse. He takes metaethics to be concerned with the theory-theory from the philosophical perspective. As we will see in the next
chapter, this interpretation is supported by the claims in the preface of the more recent JPS (2004) volume, a collection of essays by JPS called *Mind, Morality, and Explanation*: “The essays range over the philosophy of mind and action, and its connections with moral theory and the philosophy of language and explanation. We intend them to be in part an advertisement for seeing connections in philosophy, for avoiding too much specialization” (vii). Here the connection with philosophy of mind is explicit.

I say the terminology is inconsistent though because despite the preceding Smith (particularly 1994) frequently appeals to mere common sense as though it were all that we had to work with, and some of these appeals are not merely attributions to the folk, i.e., claims or observations about the commonsense conclusions of the folk. They are, rather, what we might call *philosophical* appeals to plain old common sense.

In an argument against the phenomenological conception of desire, for example, he asks whether we really believe that desires are states that have phenomenological qualities essentially. He answers: “I should say that, as least as far as commonsense opinion goes – and what else do we have to go by in formulating a philosophical conception of folk psychological states? – we evidently have no such belief” (1994: 108). I suppose this is a Moorean sort of move. Philosophy for Smith is different, if it is, only in degree and not in kind from ordinary folk discourse about very big and basic questions. It is, as he says, *merely* more systematic and technical. In this case, perhaps he really does take himself to be interested in assessing commonsense moral psychology from the folk (commonsense) perspective – a project rather different from the philosophical project connected to philosophy of mind.

Still, this is a peculiar move because philosophy *is* more systematic and technical. The folk might not think there is anything commonsensical about essential definitions,
phenomenological qualities, the phenomenological conception of conative states, nor the dispositional account of those states, all of which Smith covers in his book. Moreover, it isn’t at all obvious that they implicitly employ these concepts in everyday discourse or practice. Compare again so-called “folk physics.” All of us employ physical principles in our everyday lives (McCloskey 1983). But few claim this puts us on par with professional physicists, particularly since so many of those principles are strictly speaking false or true only in limited cases. So before considering others who engage in this dispute I want to argue directly that Smith cannot legitimately describe his project as one concerned with the folk perspective. I want to show that he should dispense with all of this misleading talk about common sense.

Hypothetically, we could make this point obvious if we could show that, as a matter of empirical fact, most of our—by which I mean the folk, though we might just as well include philosophers here too—intuitions about various philosophical questions are shaped by further philosophical commitments and issues. For example, it is probably unusual for anyone working outside the philosophy of science to have any strong pre-philosophical intuition about the best account of causation. If pushed to choose, I think most would want to know just what rides on it first. This is not to claim that the folk lack any ideas about causation or to deny that they learn how to wield the concept. But wielding the concept and analyzing the concept so as to choose between, say, casual-mechanical accounts and marked transmission accounts are two very different things. We should be careful about how much we attribute to common sense.

The same goes for analyzing the precise nature of mental states or the sufficient causes of moral motivation in commonsense “folk” moral psychology. We need more evidence for this analysis in folk discourse if we are to characterize our project as an assessment of commonsense moral psychology in commonsense terms. It does not follow from the fact that the folk talk
about moral judgments that they must have pre-philosophical intuitions about whether such judgments are necessarily or sufficiently motivating, or whether they must be necessarily or sufficiently motivating in virtue of being the kind of mental state they are. There is a substantive philosophical project involved in showing that they have such intuitions at all.

Nor does the fact that we can goad the folk into answering philosophical questions pre-philosophically necessarily tell us anything useful about folk discourse in general. For when we try to force these “folk” to reflect on some philosophical issue of interest, as in recent experimental philosophy research, e.g., Nichols 2002, it isn’t clear that we’re really uncovering genuine pre-philosophical intuitions about that particular topic—say the relationship between moral judgment and action—rather than some other. There may be evidence that folk intuitions on some (philosophical) topics are sometimes the product of considerations that are tangential, irrelevant, or external to that topic.

Knobe (2003a, 2003b) for example takes himself to have shown that folk intuitions about whether an agent’s action is intentional really depend upon evaluative considerations, like whether the outcome of the action is morally good or morally bad. It seems that subjects are more likely to say that the chairman of a corporation intentionally harmed the environment when his indifference results in environmental damage than they are to say that he intentionally helped the environment when his indifference results in a positive environmental change.

Though this surely isn’t Knobe’s point, I think that some experimental work in philosophy if it shows anything just shows that there are some topics—attributions of intentionality in conditions of vague description, for example—about which the folk just don’t
have any real pre-philosophical intuitions.\textsuperscript{28} It is just not obvious that all philosophical questions are meaningful questions when couched in less technical, everyday terms – the kinds of terms that the folk have at their disposal. Much of doing philosophy is having thought long enough about an issue to even begin to know what the interesting questions really are and why, or to have the conceptual tools to pose them.

So it may well be the systematic and technical nature of philosophy, merely or not merely, that makes these disputes possible. Why then should we appeal to plain old common sense to deal with these questions, or for that matter, why should we think common sense itself has anything to offer on these abstruse questions? We need not deny that common sense or intuition is a part or even a crucial part of conceptual analysis. But I doubt that Smith (along with many other philosophers) takes common sense or intuition to be all of conceptual analysis for the simple reason that it requires among other things some rather uncommon knowledge of linguistic phenomena and technical devices, some of which are still highly controversial in the philosophical community (e.g., the analytic/synthetic distinction). In fact in Smith’s case it involves Ramsey-Lewis sentences, which is even worse. There is a sort of meta-cognitive ascent involved in analyzing mental state concepts. And that involves knowing how to think systematically about them.

Consequently I think we should answer Smith’s rhetorical question—‘what else do we have to go by in formulating a philosophical conception of folk psychological states?’—thusly: we have philosophy to go on in formulating a philosophical conception of FP states. We have a highly systematic way of thinking about these issues and a repertoire of subtle distinctions

\textsuperscript{28} Unfortunately many experiments in so-called “experimental philosophy” probably don’t show much of anything at all. In several cases the methodology appears questionable or seriously flawed (see, for example, Weinberg, Nichols & Stich 2001). But I make no claim here about whether this is true of Knobe’s work.
developed through careful argumentation by professional scholars and philosophical giants to help. It is as Smith himself says a *philosophical* conception of folk psychological states that we are after, not a *folk* conception of folk psychological states.

If that’s right, then Smith is sometimes a moving target. At times the project is purportedly a commonsensical one, as in the previous passage. Then suddenly this talk of commonsense opinion and “mere” differences in degrees of technicality and systematizing gives way to a full-blown functionalist treatment of moral psychology and a series of complex arguments for a dispositional account of desire. I think this is a good example of how some moral philosophers have been willing to vacillate between two projects when it suits them. If you want to make some assumption in your philosophical theory seem obvious, ground it in common sense; if someone denies that there is anything commonsensical or obvious about that assumption, turn to a complex philosophical device. What gets lost in the move is the difference from speaking from *within* a perspective and speaking *about* a perspective.

Now much of what is at issue here are questions about language usage. As Frank Jackson puts it, “Although metaphysics is about what the world is like, the *questions* we ask when we do metaphysics are framed in a language, and thus we need to attend to what the users of the language mean by the words they employ to ask their questions” (1998: 30). I agree. But where Jackson’s treatment of the project really becomes illuminating is in his claims about the implications of this for the role of conceptual analysis, the bedrock of traditional philosophy. Jackson argues that what we are engaged in when we argue about interesting philosophical questions is a dispute about our *ordinary conception* of the matter:

What then are the interesting philosophical questions that we are seeking to address when we debate the existence of free action and its compatibility with determinism, or about eliminativism concerning intentional psychology? What we are seeking to address is whether free action *according to our ordinary conception*, or something suitably close to
our ordinary conception, exists and is compatible with determinism, and whether
intentional states according to our ordinary conception, or something suitably close to it,
will survive what cognitive science reveals about the operations of our brains (1998: 31).

The crucial thing to notice is the assumption that there is an ordinary conception for any one of
these issues “interesting” philosophical questions, including determinism, the nature of
intentional states, and so on. Why should we think this?

For Jackson it is because we can identify our ordinary conceptions with our intuitions
about possible cases. But this just pushes the issue back one step since it takes for granted that
there are any reasonably consistent pre-philosophical intuitions about such cases to be
discovered. Intuitions on any given question can be influenced by a host of things, including the
way in which the question is posed, the consequences we face in giving a particular answer, and
so on.

In the moral case, of course, the issues won’t be the existence of intentional states and
free will but rather the precise nature of those states, whether moral judgments are sufficient
causes of motivation, whether they motivate necessarily, whether they state facts, and so on. To
insist that there are any consistent pre-philosophical intuitions on these matters among the folk
that are in need of explaining is just to assert the folk theory-theory and consequently to insist on
an interest in speaking from within the folk perspective. It is to assert that we are interested in
taking the folk perspective on FP, which is not just a matter of the way that the folk describe,
predict, and explain behavior on an everyday, case-by-case basis, but rather a matter of engaging
with their theory containing psychological laws relating putative mental states to one another.
Contrast “Jane went to the fridge because she was thirsty” with “thirst is a mental state such that
organisms that undergo it necessarily engage in fridge-seeking behavior.” Though philosophers
might be comfortable with the idea that the latter law-like translation subsumes the former, it is
an open (and presumably empirical) question whether the folk have anything at all like this in mind.

Moreover, it is far from uncontroversial that the folk must employ such a theory (whether they do so explicitly or implicitly). This is exactly why philosophers observe the distinction between FP and the theory-theory in the first place. What we lack here is any convincing evidence in favor of the idea that either the distinction should be collapsed or that one captures folk practice better than the other.

A recent attempt by Nichols (2002) to uncover some deep-seated principles of folk theory involved presenting undergraduates with brief descriptions of fictional moral agents and asking whether that agent really understands moral claims. Here are two such descriptions.

John is a psychopathic criminal. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. John has hurt, and indeed killed, other people when he has wanted to steal their money. He says he knows that hurting others is wrong, but that he just doesn’t care if he does things that are wrong. Does John really understand that hurting others is morally wrong? (289)

Bill is a mathematician. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. Nonetheless, Bill never hurts other people simply because he thinks that it is irrational to hurt others. He thinks that any rational person would be like him and not hurt other people. Does Bill really understand that hurting others is morally wrong? (289)

According to Nichols, 85% of subjects claimed that John did in fact understand the moral claim. Moreover, Nichols claims that this cannot be due to reluctance to deny genuine moral judgment since “a majority” of subjects denied Bill understanding. In an endnote Nichols explains that a McNemar’s test revealed a statistically significant difference (p < .025) in the subjects’ responses to the psychopath and mathematician cases.

Technical/statistical questions aside, what is this experiment really capable of showing in principle? What we need is evidence that the folk implicitly or explicitly employ psychological
laws, or law-like generalizations to explain the cognition and behavior (and not anything else, such as, say, moral culpability) of others, like John and Bill. But more specifically it looks like we need evidence that the folk implicitly or explicitly subsume such cases under a law or law-like generalization (or at least somehow relate them to a general pattern). For again if we show only that we philosophers can subsume these token explanations then we have given up the commonsense project of describing the folk from the perspective of the folk.

I suspect many will want to object here that talk of subsuming under laws, variations on the DN model, etc., need not enter into the issue. But it is hard to see how we can legitimately ground such an objection. If we deny that we need to appeal to psychological generalizations or laws in a commonsense FP project at all then it not only looks like we are suddenly willing to abandon the theory-theory idea but it also becomes nearly impossible to make sense of Humeanism dispute, which is one about the general causal efficacy of a putative mental state type, i.e., a generalization about what some kinds of mental states do to people who undergo them. If the folk have any real grip on the nature of moral motivation in general, and if there is any real value in exploring the strength of that grip, then it seems like they must in some way or another be relating token instances to more abstract descriptions of behavioral patterns, and that sounds a lot like relating observed cases to a generalization or law.

So it seems to me that philosophers like Smith, Jackson and Nichols are truly committed to the law-like generalizations and for all that still frequently willing to insist upon the “common sense” perspective. And in that case the view is starting to look suspiciously like the claim that folk discourse relies upon something a lot like the DN model of explanatory understanding, which is, as Churchland (1989) has influentially argued, psychologically unrealistic in several ways. For example, it requires that we attribute to the folk knowledge of universally quantified
conditional statements, knowledge of the deduction relation, time to carry out the deduction, and lots of other things that the folk clearly just don’t have (199).

The kind of research that Nichols and Jackson cite and conduct raises additional problems for insisting on the folk perspective. As we have just seen, for whatever reason people are apparently more willing to attribute intentional action when the outcome is bad than when it is neutral or good. Nichols says that 85% of subjects attributed understanding to John, in whose case they were told people were actually killed. It could be that subjects were eager to blame John for the consequences of his actions. Fittingly, in Bill’s case—in which they were told nobody was actually harmed—they denied understanding of the moral claim, perhaps essentially denying Bill moral approbation for his self-control. In short, we might have here little more than evidence for the idea that the folk are stingy with moral praise and liberal with moral blame. Or, alternatively, support for Knobe’s idea that people are stingy with attributions of intention in cases of success and positive results and liberal with attributions of intention in cases of failure and negative results. For starters, then, we’d need to control for this potential confound just to get the research on solid methodological ground.

Moreover, even if we are willing to overlook this particular potential confound, it is still too quick to conclude as Nichols does that the folk think “the psychopath did really understand that hurting others is morally wrong despite the absence of motivation” (2002: 289; italics added for emphasis).

For one thing, the passage neither contains or implies anything clearly about motivation, inclination or the like, save for (at best) a possible tacit reference embedded in the fact that John says he doesn’t care about the fact that hurting others is wrong. The word “motivation” does not actually even appear in the description of John, so it is entirely unclear how the subjects would
know anything about John’s level of motivation, or consider his level of motivation, or assess John’s moral knowledge despite the absence of motivation. By subtly tacking this phrase on to the end of the sentence, Nichols gives us the impression that the study successfully drew out folk principles concerning the nature of the relationship between moral judgment and moral motivation in general. But that is far from clear given the methodology.

Worse still, if the question is whether we should really believe John when he says that he understands the moral claim, then his credibility must already be in doubt as far as the subjects are concerned. Why shouldn’t the subjects also think that John is dishonest about his ("motivational") indifference? The subjects are not explicitly told to take him at his word, nor are they told that he is an honest and forthright person, or anything else about his character aside from his psychopathic behavior. There are too many moving parts here. We have no way of knowing from this study on what basis the subjects were forming these views of John and Bill. That is, we don’t know whether the folk form these judgments explicitly based on any kind of law relating moral belief and motivation. Nor for that matter do we know if they do so implicitly, given Knobe’s results, which shows that there is at least one plausible competing explanation for how subjects form judgments about others.

It is worth noting that Nichols agrees with half of my claim here, which is there are some matters about which the folk lack any real, consistent intuitions. Though he seems to take it as obvious that the folk have intuitions on issues such as the nature of psychopathy and moral judgment, he agrees that those intuitions are inconsistent. In particular, he thinks there are significant differences in these intuitions across Eastern and Western cultures (Weinberg,
Nichols, & Stich 2001). The crux of my disagreement with Nichols here lies in the fact that I don’t see—and he doesn’t really say—why he thinks philosophers should continue taking putative folk intuitions on technical philosophical issues seriously despite the inconsistency.

There is another potential problem here that may be worth keeping in mind. Not only do we not know whether the folk employ (tacitly or explicitly) law-like generalizations in their explanations, we actually don’t even know whether there are any such laws to be found in philosophy. Gauker (2005) argues convincingly that so far all of the many “belief-desire laws” (that is, laws governing intentional mental states, such as those that relate beliefs to desires) that have been proposed on behalf of the theory-theory can be refuted just by thinking carefully about them, in advance of any scientific research at all. Consider, for example, a simple version of a belief-desire law that says, “people will do what they believe will satisfy their desires.” Gauker argues that this is already evidently false, for people have lots of desires and they simply cannot do everything at once (126). The proposal is already in need of modification. Gauker continues by modifying and reconsidering, ultimately canvassing an extensive list of proposals and showing how each can be refuted just by careful armchair reflection.

If he is right, we will need a good reason to think that his criticism of belief-desire laws in general does not hold for belief-desire laws in moral psychology. But far from providing any such reason, some of the philosophers working on intentional moral psychology are, as we have just seen with JPS, committed to the idea that moral psychology and philosophy of mind are closely connected. Many claim that intentional moral psychology is properly a subset of

---

29 The Weinberg, Nichols & Stich (2001) paper cited here is, in my view, seriously methodologically flawed. The authors not only conduct a study on “cultural variation” in folk intuitions by collecting data exclusively at an American university, but also at one point report a statistically insignificant result as indicative of a “pattern” in the data which supports their philosophical conclusions (18).
intentional psychology. If we take Gauker’s point, Smith, Nichols, and the like end up committed to something even stronger than the claim that the folk are an untapped source of philosophical knowledge; they end up committed to the idea that the folk have a tacit handle on an explanatory base that philosophers, for all their training and time spent on the matter, are evidently unable to explicate adequately.

Of course, you might object by just insisting that it is obvious that the folk have consistent intuitions about these questions—what I’m calling philosophical questions—and ridiculous to claim otherwise. That is, we don’t need experimental evidence that the folk actually employ law-like generalizations about cognition and behavior or the functions of mental states to explain particular cases because such evidence is all around us. This seems to be the idea that inspires many of the arguments given in defense of folk psychology so I will respond to this objection more thoroughly in the next chapter. For now I offer two simple replies.

First, then it seems you have already parted ways at least to some degree with Jackson (and presumably Smith and Pettit to the extent that they insist on a collective project) for Jackson himself admits that folk conceptual analysis and intuition of the sort needed here is a work in progress:

A person’s first-up response as to whether something counts as a $K$ may well need to be discounted. One or more of: the theoretical role they give $K$-hood, evidence concerning other cases they count as instances of $K$, signs of confused thinking on their part, cases where their classification is, on examination, a derivative one… their readiness to back off under questioning, and the like, can justify rejecting a subject’s first-up classifications as revealing their concept of $K$-hood (35).

Jackson’s point seems to be that if we want to get at the core of the folk intuition on philosophical matters we have to do some on-the-job training. But I don’t see how this is any

---

30 I think this passage reveals a tension in Jackson’s position. The claims made in this passage suggest that we might have to do some work to extract the kinds of consistent intuitions needed
kind of defense of commonsense intuition at all. It is not clear how much value there is in a pre-
philosophical intuition once that intuition is shaped by questioning or otherwise tainted by
something approaching cursory philosophical training. It was the folk’s attributions that Smith
said he wanted to explain, not the philosopher’s. So it seems peculiar to claim that in many cases
we may simply need to discount first-up responses.

And what do the folk say about the Humeanism dispute itself? The studies have not been
done. But here is a reason—my second reply to the objection—to doubt that they are worth
doing anyway. As we have already seen, Smith opens his book with a discussion of some
intuitions about the nature of moral judgments. On the one hand, they have objective purport; on
the other, they are fundamentally practical. Given that the Humeanism dispute is ultimately one
about which of these more basic assumptions to privilege—notice that if you really think moral
judgments are supposed to be fact-stating then you will probably deny the sufficiency claim and
hence accept Humeanism—the interesting question for the folk is not about Humeanism as a
theory of moral motivation but about their conception of the kind of mental state a moral
judgment is supposed to be. The philosophical implications just fall out of the answer.

Now it seems we must ask: what do the folk actually say about the kinds of mental states
that judgments are? Again, we don’t know. But it is probably inconsequential. For unless the
folk give us the Humean story about the distinction between beliefs and desires while answering
that moral judgments are beliefs, the Humeanism dispute about motivation never even arises.
This we saw in our discussion of the moral problem. It is the commitment to beliefs and desires
to adequately defend the claim that the folk are a lot like philosophers. Yet in the article that he
co-authors with Pettit (“In Defense of Folk Psychology”) the point seems to be that the folk have
to have these intuitions and that they are constantly putting them to use. It certainly isn’t a
contradiction, but I find it difficult to reconcile the latter claim with the idea that we will struggle
to extract such intuitions without some cursory philosophical training. It’s hard to conceive of a
view of folk discourse that could easily explain both of these claims simultaneously.
as distinct states that makes each of the other substantive philosophical disputes possible. In other words, the objection might work for FP in general but it runs into trouble in the context of moral motivation. The only possible set of circumstances under which we could coherently construe the dispute about the Humean theory of motivation as one about how best to describe folk explanations of folk practice is if we could actually show that the folk actually rely upon the Humean conception of mental states as philosophically construed in their everyday lives.

But even if the folk do in fact have implicit knowledge of the nature of mental states, a claim I think we have reason to doubt given the well-documented difficulties we encounter in trying to make that knowledge explicit, we have been given few reasons to think their implicit knowledge matches a substantive 18th Century theory of the nature of mental states, one on which a crucial difference between beliefs and desires is that only the latter are rationally unassailable. Is this really in the background when the folk say that Jane went to the fridge because she wanted a beer? It is possible, of course, but highly improbable. Philosophers are surely not in agreement about the distinctness of beliefs and desires, so there is no reason to think the folk have established that agreement implicitly and spread the knowledge as lore.

Worse still, even if philosophers agree with Smith that the distinctive mark of desires is that they are “not subject to any sort of rational criticism at all,” several observations indicate that the folk may actually disagree (8). Take for example the retributivist nature of the U.S. legal system (see Greene & Cohen 2002 for example), or the popularity of Dateline NBC’s show To Catch a Predator in which would-be pedophiles are publicly embarrassed and subsequently arrested for attempting to satisfy their sexual desires to the delight of the viewing audience and a (ironically creepy) television personality. The fact that so many people think that pedophiles are wrong to have the desires that they do, or that punishment ought to be about getting revenge on
criminals who act on (or try to act on) their desires without regard for the well being of society, suggests that it is at least plausible to suppose that most of us think desires are indeed rationally assailable. For many people, there are some things it is just wrong to desire. Plain observations like these actually look like prima facie reason to hypothesize that the folk do not implicitly employ Smith’s careful distinction or the 18th Century theory of psychology that we supposedly get from Hume. At the very least these observations give us some reason to hesitate. So we don’t know whether the folk really know and do all of these things when they formulate their claims about Jane’s fridge-seeking behavior. But in light of these considerations, it seems risky to hang an entire philosophical project on the claim.

This brings me to a related point about Smith’s first chapter. There is one question on which Smith is unclear and vague. That is the following question: is the so-called “standard picture of human psychology” (according to which beliefs and desires are fundamentally distinct existences) Humean or commonsensical? First Smith says that “we seem to think that moral questions have correct answers” and that moral judgments seem to be “opinions about the reasons we have for behaving in certain ways” (7). But then he tells us that these two distinctive features of morality have opposing implications, which we can see once we “reflect more

---

31 One could reply, quite plausibly, that in fact people don’t blame sexual predators for their desires but only for acting upon them. This is a difficult empirical question, and I acknowledge that this is a plausible interpretation. Still, what I take to be most telling about the To Catch a Predator Case is the very public nature of the intervention and the subsequent humiliation to which the would-be pedophiles are subjected. There is something very retributive about the act of televising people as they attempt to satisfy their desires even as they express awareness to the show’s host that what they are trying to do is wrong. Presumably if most of us thought that such people were mentally ill and in need of treatment for their desires rather than guilty and abhorrent then we would advocate treating them as we treat others who suffer involuntarily from mental illness, that is, a bit less cruelly. And in that case we should expect a bigger public outcry about televising such a crude and uncompassionate form of intervention. That there is no such outcry suggests that it is at least plausible that people really do blame others for their desires and not just for acting (or trying to act) on them.
generally on the nature of human psychology.” And this is where he outlines the standard picture, which we “owe to Hume” (7).

But in thinking through the features of moral judgment it seems to me that we have already reflected generally on the nature of human psychology. Perhaps it is true that all of the folk (presumably the “we” in the above passage) might agree with us when we explain how we can conceive of moral judgments in just these two ways. But the interesting question from the standpoint of describing the folk is whether they arrive at just these two features of moral judgment on their own, without any setup or goading. There is plenty of evidence to show that we can get people to assent to a lot of ideas by presenting them in the right way. For example: “Doesn’t it just seem like moral judgments state facts since we disagree so much about them?” “Doesn’t it seem like your moral judgments are really just expressions of your emotions since you feel them before you think them?” The folk picture as presented here—the intuitions that “we” have about the kind of thing a moral judgment is—is really just a substantive theory of 18th C. philosophical psychology in disguise. And now it might be clearer why this basic question goes unanswered in an otherwise clear chapter. There is a deep ambiguity about the nature of Smith’s project, i.e. about the perspective it assumes, and that ambiguity makes it very difficult to advance a methodological objection against him.

I think all of this suggests not just that philosophers have vacillated freely between two projects as it suits them but also that a legitimately commonsense conception of the Humeanism dispute is potentially incoherent and in any case closed to any real solution. A solution will be possible only if the folk actually think in the kind of terms in which Humeanism and anti-Humeanism are couched. We haven’t been given any compelling reason to think that they do, and we have just seen several reasons to be skeptical. There are some questions about which
people have relatively unaided intuitions and some about which they do not. Several of the “interesting” philosophical questions, to put it as Jackson does, smuggle in a lot of philosophical baggage.

So it is doubtful that most projects, including those of Smith and Jackson, amount to literally describing folk practice from the folk perspective. Little (1997) seems to agree. She is fairly clear that she takes disputes about FP mental states, moral agency and moral knowledge to be deeply philosophical: “What I want to argue is that, if we are wary of [besires], it will not be in virtue of their conceptual oddity or because of plain facts of moral psychology; it will be in virtue of further substantive disagreements on familiarly contentious issues outside of mind and psychology” (1997: 61). Her opponent in this is Smith. What we have so far said about Smith’s emphasis on common sense may help to explain why Little rejects Smith’s idea that besires are just conceptually bizarre.

Blackburn (1998) like Little also seems to think that the dispute is deeply embedded in the philosophy of mind and action. He begins the preface thusly: “…Anscombe once said, rightly, that there could be no philosophy of ethics that was not founded on a proper philosophy of mind” (v). In the fourth chapter of the same book, Blackburn criticizes cognitivism—interpreted as the view that the task of moral philosophy is to give an account of the content of ethical propositions—by showing how several recent attempts from cognitivists to skirt Moore’s open question fail. Two of those recent attempts are related and come from the JPS program. In Blackburn’s criticism there is a line of argument targeted at cognitivists very much like the one I have so far been presenting here against Humeans and anti-Humeans.

According to Blackburn, philosophers like JPS are in the business of giving “response-dependent” accounts of ethical terms. They want to analyze ethical concepts in terms of actual
human responses, such as attitudes, desires, and so on. For example, a response-dependent account might claim that X is good iff people are disposed to [insert response] under [insert circumstances]. They then fill in the brackets by specifying who exactly is under consideration (the folk, experts, etc.), what kind of reaction (cognitive, non-cognitive, judgment, etc.) will fit the theory, and the circumstances under which the reaction is supposed to occur (ordinary circumstances, conditions of paying careful attention to X, etc.). We will return to the details of the JPS program in the next chapter. Here I just want to note that in laying out his argument Blackburn makes clear which project he takes to be at issue, and he does so in a way that I think mirrors (to some degree) my complaint in this chapter:

I shall summarize my objection in advance: either [response-dependent accounts] give us no real theory of the ethical proposition, but simply substitute one way of putting it for another. Or they confuse speaking from within a moral perspective with describing those who speak from within it… To proceed, we need to notice an important division between two projects. Are we trying to understand the nature of evaluative judgment? Or are we simply concerned to ‘give a truth condition’ for the evaluative predicates, ‘X is good/right/justifiable’ (107)?

It is the sentence beginning with “or” that interests me in particular here. Blackburn’s argument is directed at cognitivism, or the idea that ethical propositions have purely descriptive content. But the strategy appears to be similar to the one I have so far been employing against commonsense moral psychology in that it draws a distinction between speaking as folk and describing the folk from the standpoint of philosophy.

Perhaps unsurprisingly Blackburn then says: “I shall eventually argue that there is something seriously wrong about seeing ethics as, at bottom, self-description” (106). This may be a variation on the idea that there is a problem about the commonsense project under consideration. On this much I agree with Blackburn, which is why I conclude as he does that the moral philosopher’s project is not properly one about commonsense FP but rather a
philosophical one rooted in the tradition of conceptual analysis and concerned primarily with getting at the philosophical Truth. Or as Blackburn says, it is an approach that “is a close descendent of classical philosophical analysis, aiming to reveal the real structure of our thought by expressing it more perspicuously, laying bare its real meaning” (87). Among the most prominent versions of this approach in moral psychology at present is functionalist moral psychology, and in particular Canberra-Plan functionalism.
3. Functionalist Moral Psychology and the Humeanism Dispute

§1. Preliminaries

In the last chapter I argued that there are problems with claiming, as Smith and others sometimes do, that the Humean dispute, and perhaps philosophical moral psychology in general, is a matter of describing what the folk do from the folk perspective. The dispute properly concerns how best to understand and systematize folk concepts and explanations from within a philosophical framework. In this chapter I complete the first horn of the dilemma outlined in the previous chapter by turning to the intersection of moral psychology and philosophy of mind.

I begin with the uncontroversial observation that many of the prominent arguments for Humeanism or anti-Humeanism are situated within the framework of the functionalist theory-theory. Several are committed to the philosophical methodology of the so-called “Canberra Plan,” a unique approach to engaging in the metaphysics of mind which grew out of collaborative work at Australian National University in Canberra. Though there is nothing novel in the idea that moral philosophers have background commitments in the philosophy of mind, I think that too many of them ignore the implications of those commitments in their moral psychology or violate them along the way. This is especially a problem for the Canberra Planners, Frank Jackson, Philip Pettit, and Michael Smith in particular, who have recently been at the forefront of the moral motivation literature. Though these Planners are explicit that a priori conceptual analysis must precede our concern with the science of human cognition and behavior, their conceptual analysis sometimes looks much more like empirical psychology conducted from the armchair. Consequently I think their methods betray the need for a move toward methodological naturalism in philosophical psychology and metaethics.
§2. Functionalism: From Mind to Moral Psychology

The idea that the goal of the Humeanism dispute is to develop a philosophical account of moral motivation using commonsense or folk intentional states is peculiar to me. Unless philosophers are concerned with the project of simply describing folk theories (of the implicit or explicit sort) of moral motivation from within the folk perspective of moral agency—what Blackburn (1998) might call speaking from within a moral perspective—it is not at all obvious why they should have a philosophical investment in the mental state types reportedly present in folk discourse. This amounts to saving roughly half of what we get from the folk, namely the terms or concepts, and substantially altering the rest by ‘systematizing’ it. The division seems perfectly arbitrary. Why should we think folk theory unsystematic and in need of philosophical help but folk state types ready for adoption into philosophical psychology? The question is straightforward enough but it seems to me that answers have been in short supply.

The answer that I have so far been defending is that philosophers revamp the folk analysis of the nature of moral agency because folk positions on the issues that philosophers care about—say, for example, the modal status of claims about the causal efficacy of a putative mental state—are sparse and in some cases nonexistent. So we preserve the terms which purportedly pick out commonsense mental state concepts, if we really do, because that is as much as we get from the folk.

In fact I doubt that philosophical moral psychology preserves much of anything from folk discourse save for some loose terminology, for there is as of yet no convincing reason to think that the folk actually use intentional mental state terms in the systematic way that some moral philosophers claim that they do, or so I argued in the last chapter. Be that as it may, philosophy has to start somewhere, so it starts with everyday terms, whether or not this leads us to assume
that such terms have a more technical role in folk discourse than they really do. That is, regardless of whether they are really part of a folk theory (i.e. the theory-theory).

We then try to invoke these commonsensical terms in technical philosophical theories, show that it works, and vindicate the idea that the folk have genuine wisdom, or “know a number of platitudes about morality,” as Nichols says (286). If it turns out that we can locate folk concepts in a mature theory of moral agency, and in particular in a mature cognitive science of moral agency, then we will have vindicated the possibility of arriving at substantive truths about the nature of moral agency from the armchair.  

This trend in moral philosophy has much in common with functionalism in the philosophy of mind. Functionalists hope to avoid scientism in the domain of the mental by showing that we can generate respectable theories of mind by looking to plain facts about what mental properties do and how people talk about what mental properties do. I don’t mean to suggest here that all moral philosophers are card-carrying functionalists, or for that matter that they are all functionalists of the same stripe. But even those who aren’t often willingly play by functionalist rules when they argue about the status of Humeanism and internalism. Most accept controversial assumptions about the nature of mental states, particularly the types of mental

32 It is not entirely clear what motivates this interest in defending the value of armchair psychology. One possibility, which I suspect, is that scientific advances are often seen as a threat to the value of traditional philosophical inquiry. Preserving the possibility of arriving at substantive truths about moral cognition and morality from the armchair would then help to prevent the encroachment of science onto traditional philosophical territory. If so, I think this is an unfortunate reaction to the advances of science. We must recognize that important philosophical questions will continue to arise as we learn more about the nature of cognitive processes. The eventual empirical irrelevance likely to result from the unwillingness of philosophers to concede authority on difficult empirical questions is a far more serious threat to philosophy than the progress of science, which is, as I see it, much less a threat than a welcome catalyst for new lines of philosophical research.
states that moral agents must have, in order to engage in moral psychology. So I need to say a bit about the ideas that underlie functionalism.

The term functionalism has come to be applied to all sorts of positions in philosophy and psychology. For our purposes we can focus on the conjunction of two central ideas. The first is that we can offer functional definitions of theoretical (unobserved) terms. The second is that we can use these functional definitions in developing a philosophy of mind. Rather than try to define mental states in terms of observable features of the world, we can define them implicitly in terms of their role in a broader network of theoretical and observable terms in a theory. Mental states are defined by what they do, not by what they are.

To take a simplistic example, we might define pain not as the firing of a particular kind of nerve fiber in the nervous system but rather as the state that is caused by tissue damage and that causes wincing, moaning, and the belief that there is something wrong with the body. David Lewis (1972) tried to show that far from being mysterious this technique is perfectly familiar. Tell any story, say your favorite detective story, by simply changing the names of the characters to variables and now you have something like implicit functional definitions of those character terms. We come to know the meanings of terms by examining the role that they play in a theory.

Pace Lewis, though, there is something a bit mysterious about the theory, namely that it seems vague or incomplete. For any functional definition (of anything) we might pose a cheap variation on Moore’s open question argument by replying, “I understand what X does, but what is X?” Common sense suggests functional definitions give us only half the story. Though it might be true that good definitions tell us about the function of the definiendum, it is not obvious that that is all they need to tell us about. Still, functionalists persist because they think that their
position is capable of avoiding the pitfalls to which other theories of mind are prone. This is true of materialism, and physicalism in particular.

Physicalism in philosophy of mind is the view that all mental entities and properties are physical entities and properties. Like naturalism in ethics, physicalism has been the dominant position in part because of the progress of psychology, cognitive science and neuroscience. Any form of non-physicalism, such as dualism or epiphenomenalism (or property dualism), will not only struggle to explain why we have made as much progress as we have in demystifying at least some characteristics or capacities of the mind via these disciplines, but will struggle to explain the interaction between physical and non-physical entities and properties.

Physicalism in philosophy of mind, like naturalism in ethics, is an ontological commitment. The idea that all entities and properties are somehow physical, or somehow “supervene” on the physical, will be plausible only insofar as we have reason to think that the tough cases like belief, desire, qualitative experience, mental causation and so on will eventually be cracked by careful empirical investigation. In what sense could my experience of redness be physical? Providing answers to such questions has been notoriously difficult.

One of the main reasons for this is the difficulty in explaining the relationship between the physics and the special sciences like psychology. Psychology trades in all sorts of concepts whose physical nature remains unknown, that is, concepts for which we have yet to identify physical “realizers.” Psychologists readily talk about beliefs, desires, qualitative experience, and so on. If everything is physical, then either these terms have physical realizers or else they are fictions. If they are fictions, then it is hard to explain why they have proved so useful to, among many others, scientists like psychologists. If they have physical realizers then they must ultimately find a place in clearly physicalist scientific theories, like neurophysiology. But where
in the wrinkled tissue of the brain do we find belief, or the brilliance of the experience of redness?

One way to answer this series of challenges is to accept physicalism, insist on physical realizers for intentional states in neurophysiology, and explain the relationship between the discipline that employs intentional states and the discipline that does not. The general position is called reductionism (or reductivism or type identity theory). Reductionists claim that the terms of the special sciences will ultimately be reducible to the terms of the physical sciences. This, they think, is why mental state types (e.g., “belief”) can be reduced to or identified with physical state types (i.e., general patterns or states of nervous system activation) to solve the mind-body problem.

Traditionally reductionists have explained the relationship between the special sciences and the physical sciences in terms of a particular model of explanation called the deductive-nomological model of explanation (hereafter “DN model”) developed by the logical positivists (Hempel & Oppenheim 1948; Nagel 1961). On the DN model, explanations take the form of valid deductive arguments. A sentence describing the phenomenon to be explained (the explanandum) gets deduced from a set of sentences, the premises, which are adduced to account for the phenomenon (explanans). The explanandum may be a phenomenon or even a scientific law. The premises contain information about general scientific laws and the initial conditions (or background information about what we already know). Here is an example from Rottschaefer (1998: 157), though most textbook science examples will suffice for mere illustration. The DN model might be used to explain why a particular piece of copper has expanded as follows:

1. All copper expands when heated. (General law)
2. This piece of metal is copper. (Initial condition)
3. This piece of metal was heated. (Initial condition)

C. Therefore, this piece of metal expanded. (Explanandum)

According to traditional reductionists the DN model of explanation can be used to explain the relationship between two sciences, like neurophysiology and psychology. The reductionist seeks to explain the relationship between two theories, namely the more basic (or lower level) theory ($T_B$; in this case neurophysiology) and the reducing theory ($T_R$; in this case psychology, the higher level theory) by deducing $T_R$ from $T_B$ plus some background and additional conditions (C). On this view, the reduction is successful when the deduction is successful and when the terms of $T_R$ have been identified with the terms of $T_B$, that is, when we have successfully specified so-called bridge laws relating the different terms of the two theories.

The DN model has been heavily criticized. Critics observed that few or no cases of successful reduction in science actually fit the model, including some of philosophy’s favorite examples from the history of science like the reduction of ideal gas laws to kinetic theory or the reduction of light to electromagnetic radiation (Kitcher 1984). This observation and many others led to a series of revisions of the DN model of reduction (e.g., Hooker 1981, Schaffner 1993). For example, some showed that what is actually deduced is not $T_R$ but rather a corrected version of $T_R$, call it $T_R^*$, which is analogous to the reducing theory. Others showed that in some cases $T_R^*$ actually corrects $T_R$. And so on. For better or worse, the many criticisms are thought to have won the day and the DN project is now widely considered a failure. Though more sophisticated versions of reduction eventually emerged, it is probably accurate to say that they never really caught on (cf. Bickle 2006; Kim 1993). The idea that mental entities and properties are type-for-type identical with physical (brain) entities and properties remains an unpopular one.
Two critics of reductionism in particular, Fodor (1974) and Putnam (1967), helped to popularize a much weaker form of physicalism than the reductionist’s type identity theory, namely the *token identity theory* or token physicalism, which does not rely on the DN model of explanation. Token identity theorists are physicalists who think that particular instances, or tokens, of mental states/events/properties might ultimately be identified with particular instances of brain states/events/properties. The view is “weaker” than type identity theory in that it claims only that any given mental instance must also be a physical instance of whatever kind.

Token physicalism was widely agreed to have at least one important advantage over type identity theory: it could accommodate the possibility that any particular mental state might also be instantiated in something or someone without a human nervous system. For example, if pain is C-fiber firing (or some similar form of nociception) as the type identity theorist holds then we have already ruled out the possibility that anything with a different kind of nervous system, or anything lacking C-fibers, can have that pain. Thus we have ruled out pain for machines, aliens with different physical means of nociception, etc. The idea that mental states might be instantiated, or realized, in multiple different systems is called *multiple realizability* (hereafter “MR”).

Putnam was the first to use the idea to formulate an argument against reductionism (e.g., 1967). His insight was that the plausibility of MR could be used as a premise in arguments against type identity theory as follows. The type identity theory claims that all mental state kinds are identical to physical kinds. MR shows that all mental state kinds can be realized not by just one physical kind but by many physical kinds. If a mental state kind is realized by multiple physical kinds then the mental state kind is not identical to any particular physical kind. But type identity theory requires this, so it must be false. For example, pain cannot be identical to C-fiber...
firing because all kinds of animals and creatures, and potentially even machines and aliens, with very different nervous systems, which lack C-fibers, can experience pain. The same goes, it seems, for other proposed type-type identities. So mental states are not type identical to brain states (see Bickle 1998 for a comprehensive review of MR arguments and criticisms).

The MR argument against the type identity theory had the effect of bringing functionalism to prominence. Type identity theory faltered because it said too much. In specifying the physical states to which mental states might be reduced it left itself open to counterexample, and in particular hypothetical examples of potential instances of MR. Functionalism, in saying essentially nothing about the physical world, had no such problem. If all that is required to be a mental state is to play a particular functional role in a system, then anything that can be shown to play that role can be shown to be an instance of that mental state.

The vagueness of functional definitions of mental states had become a philosophical virtue. Never mind that the mind-body problem is fundamentally about explaining in a meaningful way the relationship between the characteristics of human subjective experience and the physical world. Functionalism offered an empirically irrefutable theory in the philosophy of mind because it left out the usual ontological commitments. The hole created by its vagueness could then be filled with a kind of promise: the things in the universe which in fact play the functional roles constitutive of mental states will probably turn out to be physical. It’s just that they need not be.

Ultimately functionalism fell from prominence in the philosophy of mind owing to a series of influential objections. Stich (1983) showed that functionalism is committed to a “narrow causal individuation” of mental states. In defining mental states in terms of their causal relations with other states, functionalism can only distinguish between types of mental states by
appealing to the causal relations into which they enter. And more specifically, such individuation is narrow: the only potential causal connections that the theory can recognize are those that obtain between mental states and other mental states, mental states and stimuli, and mental states and behavior. Any other causal connection, such as one between my current mental states and some historical events that occurred prior to my birth, are excluded, even where such connections are important.

Kim (1989, 1998) presented functionalism with the causal exclusion problem. If mental state types are ultimately physically realized and defined by their causal relationships, then insofar as there are already physical laws connecting the mental state type to behavior, the functional role properties of the state will be causally irrelevant. To take a simple example, if pain is realized by C-fiber firing, and there are physical laws linking C-fiber firing to wincing, then C-fiber firing’s being a pain is causally irrelevant.

Others, perhaps most notably Ned Block (1980), argued that there are some kinds of mental states that resist functional definition. In particular, qualia, or states characterized by the experience of a property such as the experience of redness, seem to present a particularly difficult case for functionalism. For it seems plausible that a species of organisms with absent or “inverted” qualia (e.g., organisms who see red where we see green) might nevertheless be behaviorally indistinguishable from us. Unsurprisingly most of these objections fall under a more general category of complaint according to which functional definitions are just incomplete. Functionalism is simply an insufficient account of the mind.
§3. Canberra Functionalism in Moral Psychology

In spite of all that, functionalist ideas continue to occupy a crucial part of the moral psychology literature. This is largely because the theory has been substantially revised and reworked into part of a comprehensive philosophical project ranging from metaphysics to ethics. At the center of that project are Frank Jackson, Philip Pettit, and Michael Smith. They form the core of what is called the “Canberra Plan.”

Proponents of the Canberra Plan are united, primarily, by a commitment to the so-called “Canberra two-step.” This is the philosophical method whereby we learn about concepts by carrying out a two-step process. First, we choose the term or concept that we want to understand. In the moral case it will be a moral concept, like “right” or “fair.” We then round up all of the “platitudes” about that concept, that is, the claims about that concept that reflect the way we (everyone, or the folk) use the term. These platitudes are then conjoined in order to determine the theoretical role of the concept in question. Famously, these theoretical roles are represented using “Ramsey sentences,” existentially quantified sentences that contain the information gathered from the platitudes. In the second step we go out into the world to determine what if anything plays the role specified.

The Canberra Plan was developed as a way to bring conceptual analysis and philosophical naturalism together. In line with tradition, the job of the philosopher is to analyze terms or concepts; in line with naturalism that analysis involves defining theoretical terms using existing vocabulary. That is, philosophy still analyzes concepts, but the analysis amounts to generating Ramsey sentences from existing vocabulary in order to define new terms, which become the objects of analysis. The method is naturalistic in the sense that it, much like logical
positivism, takes philosophical work to amount to the analysis of concepts or clarification of terms (Braddon-Mitchell 2009: 25).

The first step is still occupied by familiar a priori conceptual analysis. Canberra Planners generally agree that an analysis of a concept is successful iff it gives us knowledge of all and only the platitudes that—in virtue of being platitudes—endow us with mastery of the concept (Smith 1994: 31). On Smith’s account, in acquiring a concept we acquire a set of inferential and judgmental dispositions, which connect facts expressed in the terms of that concept to other kinds of facts (37-8). The platitudes about that concept are then a set of statements of all these various dispositions. Conceptual analysis is just the attempt to articulate all and only these platitudes. To analyze moral rightness is to specify which property is the property of being right by referencing descriptions of the inferential and judgmental dispositions—platitudes about rightness—given by those who have mastered the concept (39).

According to Smith, there are many such platitudes about moral rightness. Some such platitudes describe the practicality of morality, such as “people who judge that it is right to A will be motivated to A, at least to some degree, and all else being equal;” others describe the objectivity of morality, and so on. It is platitudes of this sort that our analysis of moral concepts is supposed to encapsulate.

There are some differences between functionalism as outlined above and the moral functionalism of Jackson, Pettit, and Smith. They are clear about the relationship they take their moral theory to bear to its psychological parallel. It is functionalist in the important sense that it defines theoretical terms (moral terms in ethics and intentional mental state terms in philosophy

---

33 Here I describe the Canberra Plan in Smith’s terms. There are a few minor differences between Smith on the one hand and Jackson and Pettit on the other which will perhaps emerge later. For my purpose here, which is just to summarize the general methodology, nothing is lost by lumping them together.
of mind) using the Lewis-Ramsey-style analysis, which references their functional role in a received theory, namely folk theory. But they emphasize two differences in the moral case. First, in moral functionalism it is neither necessary nor a priori knowable that what ensures the fulfillment of the roles associated with psychological properties are physical properties; rather, it is necessary and a priori knowable that what ensures the fulfillment of the moral property roles are descriptive properties. Second, unlike its psychological analog, moral functionalism does not claim that functional definitions of moral properties must be couched in terms of their causal properties. For example, an act may be right because it is fair, but that action’s being fair does not causally explain its being right (Jackson & Pettit 1995: 195). Importantly, though, these differences probably have limited applicability to the Humeanism dispute because that dispute concerns the nature of traditional mental states, namely belief and desire, rather than moral properties like rightness and fairness.

In the remainder of this chapter I offer two criticisms of functionalism in philosophical moral psychology. First, I take my argument in the previous chapter to cast some doubt on the existence of folk platitudes as the Planners construe them. It isn’t clear to me that the kinds of platitudes required by Canberra-style functionalism are out there to be had. In what follows I will say more about my reasons for this by updating the argument I gave in the previous chapter.

Second, I think there is reason to doubt that Canberra Planners are entirely true to their own methodological commitments. In particular, I doubt that the theoretical terms which become the objects of analysis are really defined by a priori analysis rather than by disguised empirical investigation. To make the point I will draw on the distinction introduced in the last chapter between the folk and philosophical perspectives. The idea that a priori analysis is capable of yielding functional roles of theoretical terms or concepts is a confusion that results
from the tendency to blend these two distinct projects. When we insist on the distinction we find
ourselves in a better position to understand why methodological naturalism is our last hope for
developing a meaningful framework for the study of moral agency in philosophy.

§4. Naturalism vs. the Canberra Plan

The methodological thesis I have been pushing in this essay has been addressed directly
by Jackson and Pettit. Actually they point out in a footnote that they were convinced to respond
to it by conversations with Sterelny and Devitt (1990: 31, n24). Even so, it seems to me that the
simplistic presentation it receives blunts its point.

The objection is, as they rightly call it, the objection from philosophical naturalism. They
mean by this ontological naturalism, though in a moment I will argue that this is in error,
regardless of whether that error is theirs or the naturalist’s. The point is properly one about
methodology. In any case, as they represent it the objection says that in order to decide what one
ought to believe in (ontologically speaking) one ought to look to science to determine what one
needs to believe in, as opposed to wasting so much time on conceptual analysis. Science, not
philosophy, will tell us about what needs analysis.

Jackson and Pettit begin by claiming that they are “sure that [the objection from
philosophical naturalism] lies behind much of the doubt naturalistically oriented philosophers
have about beliefs and desires” (31). The thrust of their response is that we have no good reason
to think that philosophical naturalism really threatens the existence of such intentional states.
There is “excellent” evidence that we are sometimes in the states which satisfy the “folk roles” of
beliefs and desires (32). To understand the evidence we have to understand what they mean by
folk roles, for when they say that we have excellent evidence that we undergo functionally defined intentional mental states they mean that we have excellent evidence for folk roles.

Jackson and Pettit claim that folk practice determines a hypothesis about how we are functionally organized. It is a hypothesis that we follow implicitly or explicitly in “moving between” behavior, situations, beliefs, and desires (16-17). The hypothesis is that we have states that fulfill these folk roles, or functional roles that the folk use implicitly or explicitly. The question as they put it is whether we have reason to accept the hypothesis as coming close enough to capturing the relevant part of our functional organization and to be confident that neuroscience won’t ultimately undermine that hypothesis.

I have so far suggested that the Humeanism project as a piece of philosophical psychology might be construed in one of two ways: as a commonsense dispute from the folk perspective and as a philosophical dispute about folk discourse. In parallel Jackson and Pettit divide the more general question of whether there exist intentional mental states into two questions: first, what needs to be the case on the folk conception for a state to be a belief or a desire; and second, is what needs to be the case on the folk conception in fact the case? In other words, the question whether there are any states which satisfy the criteria for being a belief or a desire is really made up of two issues: what are the folk criteria, and are those criteria any good philosophically? They call these the folk question and the factual question respectively.

Their motivation for drawing the distinction is to develop an argument against eliminativism that takes the following form:

---

34 Their distinction here appears to support the distinction I make in the previous chapter between the FP theory and FP theory-theory projects. One way to rephrase my complaint with Smith’s ambiguity in that chapter, then, is to say that he sometimes seems to vacillate between the folk and factual questions.
1. It is sufficient for having beliefs and desires that one be in states which satisfy…[fill in the blank].

2. We are sometimes in states which satisfy … [fill in the blank].

3. Therefore, we have beliefs and desires.

(1) is the folk question, (2) the factual question, and (3) the anti-eliminativist conclusion.

According to Jackson and Pettit, the goal is to find a way to fill in the blanks with a uniform statement that makes both premises true. The catch, however, is that the premises tend to work against one another because that statement needs to be strong enough to give us sufficient conditions for a state’s being a belief and desire in (1) and yet not so strong as to make it plausible that neuroscience will provide reason to doubt it somewhere in the future in (2). They think that they can give a statement that achieves this golden mean by invoking the notion of folk roles as contrasted with psycho-functional roles, where the latter are roles that future science will reveal (16).

The idea is to defend a version of commonsense functionalism by claiming that it is an empirical fact that people have both implicit and explicit knowledge of psychological states. We predict and explain each other’s behavior with a great deal of success, they claim, and in many cases we do so by employing belief-desire hypotheses. What make our success in this practice possible are the functional roles. We associate those roles with beliefs and desires. Importantly it is not the thing that realizes the role, but just the functional role itself upon which the folk rely.

Now on this view the objection to commonsense functionalism which claims that the folk simply don’t know enough about mental states to determine which functional role is definitive of something’s being a belief or a desire—an objection like the one I gave in the previous chapter—fails on the ground that it ignores implicit (or tacit) folk knowledge. The very fact that the folk are so successful at predicting and explaining behavior shows that they have a good deal more
knowledge about the functional roles definitive of mental states than the objection lets on. The success of the folk in the practices of prediction and explanation is made possible by the fact that they presuppose the functional roles of belief and desire.

Jackson and Pettit claim that the blank spots in (1) and (2) can now be filled as follows:

1. It is sufficient for having beliefs and desires that one be in states which satisfy the folk roles.
2. We are sometimes in states which satisfy the folk roles.
3. Therefore, we have beliefs and desires.

It is now supposed to be the case that (1) and (2) are relatively uncontroversial. (1), they argue, is both a plausible account of the folk theory of beliefs and desires and it makes no claims about the well known controversies surrounding functionalism, some of which I have just outlined in my brief historical sketch. (2) might be controversial if it is plausible that neuroscience will ultimately give us reason to doubt that we are sometimes in states that satisfy the folk roles. Perhaps neuroscience will eventually show us that we are never in such states. One good reason to think this is that we tend to assume that it is part of the commonsense view of mental states—take beliefs here—that they cause verbal and motor behaviors and consequently that the same state causes both. This idea seems a plausible candidate for rejection by future neuroscience.

Stich has registered this objection:

A belief that p is a belief-like state similar to one which we would normally express by uttering ‘p’. But, of course, the belief that p does much more than merely contribute to the causation of its own linguistic expression. It may interact with desires in many ways, some of which will ultimately issue in nonverbal behavior. It is a fundamental tenet of folk psychology that the very same state which underlies the sincere assertion of ‘p’ also may lead to a variety of nonverbal behaviors. There is, however, nothing necessary or a priori about the claim that the states underlying assertions also underlie nonverbal behavior. There are other ways to organize a cognitive system (1983: 231; italics are Stich’s).
Jackson and Pettit reply that it is no part of their view that the folk role of belief typically causes both verbal and motor behavior and that this one state causes both. This is why the notion of a folk role is crucial to their argument. The folk role is merely the idea that there is a state with such and such a role, and sometimes knowledge of that role is implicit or tacit as well as explicit. But it is no part of the implicit or explicit knowledge that the folk use in their predictions and explanations that there is some particular number of states or that a single state causes both motor and verbal behavior. The folk practice determines a hypothesis about our functional organization (24). If this is right, Stich’s objection is not a problem for their view and so the anti-eliminativist argument is sound. I’ll return to a version of this objection shortly.

This sets up their response to the objection from philosophical naturalism according to which we do better to look to science rather than conceptual analysis to determine the ontological claims needed by the sciences. Recall that the response was that we have “excellent reason” to believe that the neurophysiological states needed to explain behavior actually occupy the functional roles definitive of beliefs and desires. We are now supposed to see that the reason is that beliefs and desires qua states occupying folk roles are required for successful prediction and explanation. Here is the crucial section of the reply:

It would be a mistake to object here that all this shows is the need for the states which are beliefs and desires, rather than the need to suppose that they have the properties of being beliefs and desires. The states having the properties of being beliefs and desires is nothing more than their filling the appropriate roles, and we certainly need to suppose that they do that... The concern of the naturalists is that in order to see the need for beliefs and desires, we had to indulge in some conceptual analysis. The science per se will tell us which roles are occupied by which neurophysiological states. The fact that this constitutes having beliefs and desires required common-sense functionalism, and so some conceptual analysis. We had to say something about how we conceive of beliefs and desires. It was, however, a very ontologically modest bit of conceptual analysis. We did not use our account of the folk conception of beliefs and desires as part of an argument that completed neuroscience is incomplete, but instead as part of an argument that beliefs and desires will very likely be found within what completed neuroscience tells us. The role of our account of the folk conception of beliefs and desires was simply
to show which part of any likely complete neuroscience story is the part which says (though not in so many words) that there are beliefs and desires (32).

There are several difficulties with this response to the objection from philosophical naturalism. First, let’s return to the start, where Jackson and Pettit introduce the objection from naturalism as the objection that “lies behind much of the doubt naturalistically oriented philosophers have about beliefs and desires” (31). They appear to have the motivational claim backwards, which suggests that they have not fully appreciated the force of the problem.

It is not that naturalists—by which I mean those concerned with a genuine naturalism who are legitimately familiar with the state of the science—are motivated to reject beliefs and desires because they believe that conceptual analysis is waste of time. Rather they are motivated to think that all of the conceptual analysis is a waste of time because the relevant sciences tend to ignore the mental state-roles—that is, the roles presupposed by the folk couched in terms of theoretical mental states—that philosophers would take to be legitimate occupants of the folk roles for belief and desire. Showing this in detail will be the topic of the next (heavily empirical) chapter. Here I want to focus on the confusion about what the naturalistic stance amounts to.

As Jackson and Pettit present the objection, it seems the naturalist comes to the conclusion that using conceptual analysis to determine what the sciences need is a waste of time independently of any practical observation about what scientists are concerned with explaining. But we should all agree that that would amount to a pointless and arbitrary commitment to naturalism. Even the most ardent anti-naturalist can acknowledge that naturalism is not an arbitrary, premature commitment to a particular method. A genuinely naturalistic criticism of conceptual analysis has (or at least ought to have) its origins in actual observations about what scientists are in fact trying to do. The point is supported by the widely accepted historical observation that the naturalistic turn was inspired by progress in the neural and cognitive
sciences (cf. Copp 2007; Darwall et al. 1992). If the folk mental state-role fails to appear in the relevant scientific research in the appropriate way then there is no point in going to great lengths to preserve it philosophically – unless of course we are strictly interested in describing moral discourse from within the folk perspective. But we have already seen that in the case of philosophical psychology we are not. This is the real impetus for naturalism.

And the point is methodological rather than ontological. It is perhaps true that some versions of naturalism are primarily concerned with the ontology, or with what entities the sciences need. But to formulate the view this way is a mistake. The real issue is that scientists seek to explain particular phenomena by postulating or referring to concepts (some of which are famously unobserved, such as ion channel gates in neuroscience – more on this shortly) that serve particular roles in their theories. Whatever physical entity might fill that theoretical role, as a matter of method we move from the need to postulate something that plays a particular theoretical role in a scientific explanation to the need for philosophical analysis of the role, not vice versa. If it turns out that the role is largely irrelevant to the scientific theory, then we must either jettison the role to preserve scientific respectability, or insist that we are merely interested in what I have been calling the commonsense project and drop the claims about scientific respectability (i.e., claim only the project of describing what the folk do from within the folk perspective).

35 As I said in the introductory chapter, the point I am making here is one about the current state of neuroscience research, not about the historical development of neuroscience. This is important because one might object that it is absurd to deny that the neural and cognitive sciences could have developed without reference to something like folk mental-state roles. That may be right, but this is not an essay about the history of science. I can grant the historical point and nevertheless maintain that as things now stand those roles have been phased out of or marginalized in the relevant research programs. This is a controversial empirical claim, and I will try to defend it in the remaining chapters. Even so, it is crucial to separate the historical claim from the claim about contemporary methods.
Jackson and Pettit do neither. They emphasize the need for conceptual analysis to determine that there are in fact some things which have the properties of beliefs and desires in the scientific theory independently of (or at least prior to) looking at the scientific work itself. And this is precisely why so much conceptual analysis is a “waste of time”—to the extent that one insists that the role will turn up in the future sciences, it is a premature commitment. But Jackson and Pettit do insist that the role will turn up in “future neuroscience.” We saw previously that their project—despite how they, and Smith, sometimes describe it—is not merely concerned with describing what the folk do from within the folk perspective.

It is important to understand why the methodological claim is the basis for the ontological one. The fact that there is some role to be played in an empirical theory about human agency raises the empirical question of what thing in the universe (or, if you are a physicalist, in the physical brain) will fill that role. The naturalist is only interested in looking at the physical role fillers for roles that scientists actually need filled. Any other role is explanatorily irrelevant from the naturalistic perspective (unless, once again, you are interested strictly in analyzing the folk discourse from within the folk perspective, in which case the scientific question does not arise).

The point becomes more concrete when we take the Humeanism dispute in particular. In that dispute we are given essentially two options concerning the causal efficacy of moral beliefs qua folk roles: either those folk roles are the sufficient causes of moral motivation or they are not. There is nothing incoherent or prima facie problematic about the dispute so long as it is confined to the folk perspective (though as I argued previously I suspect in this case it will admit of no solution). But once we accept that the Humeanism dispute is a philosophical one rather than a folk one—a “factual question” rather than a “folk question” to use Jackson and Pettit’s formulation—we take on the question whether those folk roles, in this case the properties of
beliefs and desires, have causal efficacy in the production of behavior from the standpoint of empirical science. And to do that is to just assume (or insist) that those roles actually appear in the relevant scientific research in the relevant ways without looking first at the actual science.

The philosophical question can be rephrased so as to make the problematic assumption more evident: Is the folk role of moral belief (as opposed to the folk role of desire) in the neuroscience of moral judgment and motivation the sufficient cause of the motivational behavior or not? The problem is that we have been given no reason at all to accept that there is any such role in the neuroscience of moral judgment and motivation. More to the point, even if there really is, it is far from clear that it is invoked in the causal explanation for the phenomenon in question.

At present it might be enough for me to insist that it the burden of proof is on the moral functionalist to show that such roles do or must appear in the sciences, and to point out that they haven’t succeeded in showing anything like this. This suffices to make the methodological point and to show that the naturalism they target is a straw man, which is the purpose in this chapter. In later chapters we can then fill in the burden of proof argument with a demonstration that in fact no such role is present in the neuroscience of decision and motivation, at least not in the form that JPS need it to be for their argument to go through.

In other words, my empirical project will require showing that there are no “theoretical realizers” (so to speak) for the postulated roles; i.e., there are currently no theoretical concepts or entities in the contemporary neuroscience of decision and motivation which occupy the position in neuroscience research that JPS claim are or must be occupied by folk roles, and there are not likely to be anytime in the future. And if that is right, then unsurprisingly there will be no physical realizers either.
One of the peculiar features of the position I am developing is that it diverges from the naturalist objection as Jackson and Pettit describe it in that strictly speaking it does not require a defense of eliminativism about intentional mental states in general. I think this approach to the issue is typically preempted by a false dilemma. Here, for example, is Braddon-Mitchell’s description of the Canberra two-step put to work in the philosophy of mind:

Thus, for example, in the classic case of philosophy of mind we might be able to specify the belief role in a Ramsey sentence, and then the only allowed move would be to see if there is something in our scientific ontology that plays that role (or near enough), in which case we have vindicated beliefs, or else we discover there is no such thing, in which case we have discovered that eliminativism about belief is true (2009: 26).

Why should these exhaust the possibilities? Rather than argue for eliminativism, which is extremely broad, highly controversial, and notoriously difficult to defend on a large scale, I think we can actually grant for the sake of argument that something near enough to the folk roles of belief and desire do in fact appear in the contemporary scientific research and then show that the capacity in which those roles serve in that research is sufficiently different from the work that the functionalist needs them to do in order to preserve the value of their approach to moral psychology, in this case the Humeanism dispute. That is, I need only show that those roles are misplaced in the moral psychology disputes, or that their theoretical realizers within the sciences fail to fit the framework of the Humeanism dispute, and not that intentional mental states or state-roles in general have no place whatsoever in the neural and cognitive sciences in general.

Establishing this much narrower claim would be a significant result for metaethics and philosophical psychology. With all of the talk about neuroscience in general it is easy to lose sight of this point. I suspect that one of the reasons there is so much controversy and confusion about whether folk mental states will ultimately be vindicated by a “completed neuroscience” is because philosophers, and moral philosophers especially, frequently work with such a vague and
overly broad conception of ‘neuroscience.’ Some seem to have the idea that neuroscience is interested in, and headed toward, a grand, unified story of human nature. But that is philosophical work, not grant-funded laboratory science as conducted all over the world by highly specialized scientists, and we must be careful not to blend these together. The relevant question for moral functionalists is really whether neuroscientific explanations of the particular kinds of moral behavior that they are concerned with (judgment, motivation, and the like) rely upon or employ theoretical concepts that look a lot like the folk roles that they have already committed themselves to on the basis of their conceptual analysis.

So we need to be clearer about which very specific neuroscientific domain is the relevant one for focusing our search. The idea that beliefs and desires will pop up in future (or completed) neuroscience is so vague as to be nearly incoherent. Surely moral functionalists do not think that their folk roles will turn out to be a part of the kind of neuroscience that investigates proteins that bind and unzip synaptic vesicles or the molecules that underlie long-term potentiation (LTP) and synaptic transmission. The concept of ‘neuroscience in general’ is just confused.

Where Braddon-Mitchell and so many others go wrong, I think, is to assume that if beliefs and desires appear in neuroscience then eliminativism is false; or conversely to assume that if they don’t appear at all then eliminativism is true. The relevant question is not whether the terms appear in neuroscience in general, whatever that means, but whether they appear in the specific neuroscience research paradigms that pertain to the philosophical theory in question. In the case of the Humean theory, that will be the neuroscience of moral judgment and motivation.

36 Interestingly, though, these molecules and the neural mechanisms with which they are associated are already under investigation for their direct, causal contribution to cognition and specific behaviors. This is especially true in the case of the CREB molecule and memory (see Bickle 2006, for example).
The second problem with Jackson and Pettit’s reply to the naturalist is that it drastically overestimates the extent to which knowledge of mental state “platitudes” are grounded in a priori reflection on concepts rather than extensive empirical research. In making this point I directly address the earlier claim that the folk have significant knowledge of platitudes.

Recall that Jackson and Pettit reply to an objection to functionalism according to which we simply do not know enough about the functional roles occupied by beliefs and desires to actually have the commonsense lore, or body of knowledge, that their functionalism requires (16). Stephen Schiffer (1987) has objected in this way. We have seen that Jackson and Pettit’s response centers on implicit or tacit folk knowledge: the Schiffer-type objection to functionalism seems to assume that the folk knowledge must somehow be explicit knowledge of psychological states manifested in an ability to demarcate one type of state from another while in fact all we need is tacit or implicit knowledge manifested in successful prediction and explanation of behavior on a regular basis, which is made possible by implicit belief-desire hypotheses, or the putting of folk roles to use.

The crucial piece of Jackson and Pettit’s response comes in during their defense of premise (2)—the claim that we are sometimes in the states which satisfy the folk roles of belief and desire—primarily by arguing that we have excellent evidence for it in the form of folk practice. In fact, they say, we bet our lives on (2) every time we drive a car. The idea is that we must accept the hypothesis about how we are functionally organized in order to do what we do every day as successfully as we do it. If we were unable to predict the behavior of others on the road we would be unable to drive our vehicles without crashing (24). Our predictive and explanatory success is excellent evidence of the existence of functional roles – we rely on those roles to adequately predict and explain what others do.
In the last chapter I argued for something very much like the Schiffer-type objection. Roughly the idea was that it is doubtful (or at least more controversial than many are willing to admit) that the folk have enough knowledge of what it is for someone to be in the state of belief or the state of desire or any other intentional psychological state that they could use hypotheses about those states in the consistent and systematic ways that philosophers like Jackson, Pettit and Smith need them to. I want to revisit that idea here with a greater focus on the commitments of Canberra functionalism.

The key point for the Canberra Planners’ version of functionalism is that as a matter of philosophical method our (philosophical) knowledge of the implicit folk knowledge of the mental goings-on of others is not grounded in empirical or scientific experience but rather in our a priori reflection on the relevant mental state concepts of the folk. The Canberra two-step is the procedure according to which we (philosophers) find physical realizers for mental state concepts, if we actually do, only after we have given the descriptions of the types of states which we come to through armchair reflection.

Unfortunately this puts me at odds with some eliminativists, particularly Churchland, as well as with Jackson and Pettit, on the question whether folk knowledge, or lore, constitutes a theory. It is unfortunate because I find many eliminativist arguments persuasive. My disagreement with Churchland (1989) about whether FP constitutes a theory is, I think, rooted in the distinction I want to draw between the folk and philosophical perspectives. Churchland seems to assume that his argument for eliminativism about commonsense mental states requires a successful defense of the premise that FP constitutes a legitimate theory. But it seems to me that in fact it requires something less. It only requires the premise that the systematized sort of FP generated by philosophers invokes laws, supports causal explanations, adapts over time, etc.,
i.e., all the things that make it subject to empirical refutation (1989: 111). In any case I, like JPS, reject the eliminativist assumption that FP constitutes a theory.

One of the reasons that Jackson and Pettit are happy to agree with eliminativists on that point is because it sets up their argument for the anti-eliminativist conclusion by granting widespread knowledge of law-like relations. Once we acknowledge those relations it becomes a short step to showing that it is from having knowledge of those relations—much of which is implicit or tacit—that people become remarkably successful at predicting and explaining behavior. To support the idea that the folk do in fact have tacit knowledge of functional roles, Jackson and Pettit point out that the alternative is unpalatable:

Now the ability to move back and forth from behaviour in situations to belief-desire hypotheses in successfully explaining and predicting behaviour shows that we have implicitly mastered, whether or not we have always explicitly noted, the needed generalizations between the inputs in the situations, the behavioural outputs, and the beliefs and desires. The alternative is to suppose that we have arrived at our predictions by chance. But then the success of our predictions is also chance—and that is incredible (17).

But there is just no compelling reason to think that success by chance must be the one alternative to the use of belief-desire hypotheses for successful folk practice. Folk roles as Jackson and Pettit construe them are hardly the only viable option.

Here, for example, is one conceivable alternative. Perhaps the hypotheses are merely a way of explicitly noting from an external (philosophical) perspective what the folk are doing. But perhaps it turns out that these hypotheses are themselves just a (useful) philosophical fiction. What the folk do when they predict and explain behavior is to make use of something like procedural knowledge: they pick up on subtle verbal, environmental, and behavioral cues which activate recall of familiar experiences of the past, similar to how they use instinctive reactions in playing sports and musical instruments, learning to walk, navigating streets without knowledge...
of street names, calming rowdy animals and children, riding a bike, and so on. Learning occurs over extensive experience with what people do in which particular environments, and instinctively we come to “expect” (or predict) particular behaviors in particular situations without relying upon any belief-desire hypotheses at all. We gradually come to have more fluid motor responses to how we feel and respond. Scientists might even be able to tell some physicalist story about how this is made possible by brain tissue: some kind of Hebbian process takes place so that when we encounter a previously-experienced situation, the relevant synaptic connections are primed and the relevant neurons communicate more readily than before.

Now, to revisit Jackson and Pettit’s automotive example, the fact that we willingly stake our lives on our predictive success when we drive our cars may be evidence that we know we are generally successful, but it really tells us nothing of why we are successful. We dread being in the car or on the road with new drivers precisely because those drivers have a limited experiential repertoire from which to draw and in accordance with which they might activate the necessary instincts (swerve, brake, etc.) in a timely manner. This is what makes them dangerous and more likely to crash than experienced drivers. The fact that we can characterize all of this from an external philosophical perspective in terms of belief-desire hypotheses, or reliance on the rather specific folk roles of belief and desire as some kind of platitudinous knowledge, is irrelevant. The real issue is whether we must.

Never mind whether anyone really thinks that the account just given is an accurate explanation for folk predictive and explanatory success. The point is that it clearly is a coherent (even plausible) alternative to the idea that such success is made possible by “mastery” of implicit belief-desire hypotheses, and it is an alternative that is not at all like success by pure chance.
Methodologically, Jackson and Pettit claim that the platitudes constitutive of knowledge of folk roles—platitudes concerning belief-desire hypotheses, or folk roles, upon which we rely—are known, or grounded, a priori. So long as we accept their dubious idea that the only alternative to implicit mastery is pure chance, there is no obvious problem with claiming that these platitudes—by which I mean platitudes having the characteristics that JPS assume they do—are known a priori. But once we see that there are (any) alternative explanations available for the success of folk prediction and explanation, it becomes clear that which is the correct explanation cannot simply be ruled out by armchair reflection. It is an empirical question requiring extensive investigation of cognition and behavior.

The conflation of folk and philosophical perspectives is, I think, what makes this move possible. Jackson and Pettit’s assumption about the lack of alternatives is really rooted in the folk perspective. That is, it is a commonsense fact that the folk trade in beliefs and desires, and that makes it obvious that either they are successful because they have mastered belief-desire hypotheses or else they are successful by chance. The trick is that we’re not supposed to challenge this claim because it is commonsensical: it is true that the folk trade and beliefs and desires, and it is also true that they are successful. It’s hard to argue with either point.

But when we move to the less practical, theoretical/philosophical perspective and thereby avail ourselves of resources that the folk might lack, particularly considerations from empirical science, all kinds of alternatives begin to emerge. Which of the two theories given above (mine or Jackson and Pettit’s) is the right one is a philosophical and empirical issue about human psychology, not a question that the folk need to settle to continue on predicting and explaining successfully. Their practice just is successful, whichever of these turns out to be the correct explanation for that success.
That the practice is successful is obvious. It might even be obvious that the practice is successful because the folk have access to platitudes of some kind or another. But this is as far as we can hope to get. Why more precisely it is successful is a substantive empirical and philosophical issue. That is, how the nature of the platitudinous knowledge makes folk practice successful is a substantive empirical question, and it is exactly the question that Jackson and Pettit’s folk roles are intended to answer.

Once we sort out the confusion about the perspective from which we are working, we see that we are interested in the philosophical (why and how) questions. And now it looks like Jackson and Pettit’s claim that we philosophers have knowledge of the nature of the folk platitudes a priori is a very strange one. How could we possibly? For if folk platitudinous knowledge turns out to be instinctual in the way I just described, for example, then even our external philosophical descriptions of the practice as given in terms of belief-desire hypotheses would be (useful) fictions and so would be strictly speaking false. In other words, philosophers might just have the nature of the so-called platitudes all wrong when it comes to the facts about how the folk really use them, even if they make sense of folk discourse philosophically. It will take quite a bit of empirical research to know whether this is actually the case.

The Canberra two-step is a philosophical procedure, not a folk one. The first step is that we list the platitudes, which is an a priori process. But though the knowledge that the folk use when they employ what philosophers call the platitudes (these might be instincts or functional roles in the form of belief-desire hypotheses) may well be of an a priori sort, it is not at all a priori knowable what those platitudes are actually like or how they enable the folk to do what they do (i.e., which platitudes as philosophically described they in fact use). The only thing we philosophers know a priori is that the folk must do something in order to be so successful at
predicting and explaining, and further that that “something” is unlikely to be luck or chance. Beyond that, nothing can be known a priori about what is actually done. To even try to list plausible candidates for the platitudes—to spell out examples of folk platitudes or to specify the nature of those platitudes—is already a substantive empirical thesis – a thesis about what the platitudes must be like that enable successful prediction and explanation.

Successful folk practice does not provide “excellent evidence” that we are sometimes in the states which occupy the folk roles of belief and desire; rather, it provides excellent evidence that we are sometimes in states of whatever kind that are sufficient to make folk prediction and explanation successful. The latter is obvious and a priori while the former is a controversial empirical hypothesis.

Jackson and Pettit might reply here that knowledge of the functional roles of belief and desire is just too obvious and widespread to be coherently denied. Consider what they say about the relationship between postulating entities and folk roles: “we are all familiar with the idea that postulating entities of kind K may be extremely fruitful for predictive purposes, and yet there are, or well might be, no Ks” (25). That is, we know that the functional roles are in fact filled; it is just that we may not know or may be wrong about what fills them. For example: “Ptolemy’s epicycles were, and a system of particles’ centre of gravity is, predictively useful postulates, and yet neither epicycles nor centres of gravity exist” (25).

I think that the subtle mistake, present in this passage too, is to assume that because it is obvious that there is some function, Y, needed to bring about phenomenon Z that we must know how Y brings about Z. But prior to scientific investigation we may know almost nothing about the nature of Y, or about what function is required to bring about Z. Jackson and Pettit seem to think that in the case of intentional psychology we know the nature of Y on the basis of Z – Y is
the (singular) function that needs to be filled. To insist that we need beliefs and desires is to insist on some fact about the nature of Y: something has to do what beliefs and desires do (or something has to have the properties of beliefs and desires).

If so, their argument that we don’t need to know anything about what entity brings about Y is a red herring. Of course they are not claiming here that beliefs and desires are physical entities of some kind; they are claiming that they are functions without which we can’t make sense of successful prediction and explanation. We were always in agreement about the former, i.e., that it is an open question what fills the roles (which makes it odd that they bother to state explicitly that the functional roles must in fact be filled). It is the latter, i.e. that the functional roles exactly as they conceive them are the roles to be filled, that isn’t at all obvious. And even if it should turn out to be true, it couldn’t have been known a priori.

Now it is true that in some scientific cases the researchers know (or knew) what Y has to be like on the basis of the nature of the phenomenon, Z, of interest. Even so, the analogy with intentional psychology does not hold. There are many different ways in which the folk could have platitudinous knowledge—many different kinds of platitudinous knowledge—that might be sufficient to bring about successful prediction and explanation: intuitions (Y₁), propositional knowledge and deduction (Y₂), some as yet unidentified sensory modality (Y₃), and so on. And, of course, any substance could equally realize any of these. But that is a different question anyway, which is why, it seems to me, it must be a red herring.

To use an example from Smith, I may fail the philosophical test for platitudinous knowledge of the concept ‘white’ by mistakenly agreeing to the coherence of transparent matte white, or thinking it possible to put such a concept to use, and all the same never have trouble in practice by virtue of the fact that I have learned to instinctively avoid mistaken attributions of
whiteness. It might even be true that I come to treat the idea that there is no such thing as transparent matte white as platitudinous but for the wrong reasons: I just never encounter situations in which references to transparent matte white make any difference to successful prediction and explanation, or my instinct to avoid such attributions overrides potential instances in which I might use it. And it is a legitimate empirical question whether this holds for beliefs and desires.

Jackson and Pettit’s point, I take it, is that hypotheses which rely upon purely functional roles, whether about mental states or epicycles, are successful precisely because something fills those roles. Indeed Jackson and Pettit conclude the analogy by claiming that if Ptolemy had given the purely functional description (rather than specifying epicycles) he would not have been refuted by Copernicus and he would have been saying something entirely obvious, just as in the case of beliefs and desires (26). But this misses the point entirely. The problem is that for all we know it might well have been that Ptolemy didn’t give a purely functional description in the first place because the precise functional nature of the role that needed to be filled, Y, was not at all obvious. In real empirical theories it will in many cases be far from obvious what the nature of the function to be filled really is.

One of the clearest examples of this, which comes from the history of neuroscience, is the discovery of the mechanism of the action potential. Carl Craver (2007) gives an account of the discovery of ion-specific channels in Explaining the Brain that I think illustrates the point clearly. In 1952 the physiologists Hodgkin and Huxley developed a mathematical model of the action potential based on several of its known features, such as the form, amplitude, and threshold, which had been established by previous research. The model was, however, just a sketch of the mechanism by which action potentials are produced because it contained filler
terms, i.e., stand-ins for some poorly understood key components of the story. They knew that action potentials were produced by the depolarization of a cell body, and that this involved the movement of ions across a cell membrane, for example, but they did not know the mechanisms through which ion movement and depolarization were possible. Here is Craver:

[Hodgkin and Huxley] consider the possibility that ions are conveyed across the membrane by active transport. They suggest that perhaps a number of “activation” particles could weaken the integrity of the membrane. They hint at a biological interpretation of their model according to which activation and inactivation particles move around in the membrane and somehow change the membrane’s resistance. They admit, however, that they have no evidence favoring their model of other possible models. This admission spurred research on the biophysics of the membrane and the search for ion channels. Nonetheless, well into the 1970s most neuroscientists regarded the talk of ion-specific channels as mere metaphor at best and boxology [essentially re-description in vague and non-explanatory terms] at worst (2007: 115).

What was crucial for explaining the mechanism of the action potential was not simply specifying what physical realizers filled the role of “activation particle” or “active transport” etc. Rather, it was a matter of pinning down how actually—by which function given the many possibilities—ions cross a cell membrane to initiate depolarization of a cell. Claiming that conceptual analysis will elucidate the cognitive functions underlying successful folk practice (i.e. beliefs and desires as folk roles) while future neuroscience will merely reveal the realizers of those functions is a lot like claiming that these scientists could have determined the existence of ion channels from the armchair, doing the biophysics only to reveal ion channel realizers. But that is false. The science revealed the function, not just the physical realizer. Alternatively, we might interpret JPS as claiming that the relevant functional role in this case is just the depolarizer. But the question is not what thing serves as the depolarizer but how a depolarizer actually depolarizes. Depolarization in this case is the product of the function, not the function itself.

The first premise in JPS’s argument says that “it is sufficient for having beliefs and desires that one be in states which satisfy the folk roles” (18). But that is not quite enough. It is
sufficient for having beliefs and desires that one be in states which satisfy the folk roles in such and such a way, i.e., in whatever manner is supposed to be characteristic one’s undergoing beliefs and desires. How exactly do belief-desire hypotheses work? It may be obvious that we are sometimes in states which facilitate successful folk practice (i.e., folk roles), but that is not the same as having evidence for beliefs and desires, which serve to specify how and not merely that folk practice is successful.

Moreover—and this is the more important problem—they need this to be true a priori, otherwise they have merely arrived at their conclusion on the basis of having given evidence for a purely empirical hypothesis about how people are able to do what they do as successfully as they do it. (1) conflates the folk and factual projects by injecting a posteriori empirical considerations about the nature of the functional roles to be filled into the a priori and obvious claim that something must make it possible for the folk to do what they do with a remarkable degree of success.

It turns out we can test this conclusion directly. If I am right, we should expect that the purported first stage in the Canberra two-step, which is a priori in theory, will actually be filled with substantive empirical claims and scientific work in disguise when put into practice. For how else would Jackson and Pettit come to assume that their particular philosophical hypothesis about which platitudes the folk need (namely belief-desire hypotheses qua functional roles) is the right one? Surely they had to look at several cases, i.e. do some empirical work, and draw a conclusion. They could not have known the nature of the functional role a priori because there are several plausible alternative accounts of the nature of the so-called platitudes.

In fact I think that when we look more carefully there really is substantial evidence that the first stage of the Canberra two-step in actual practice often turns out to be shoddy armchair
psychology disguised as platitudes analysis. If so, then the moral functionalists have violated their own commitments in the philosophy of mind. I now turn to some of that evidence.

§5. Canberra-Style Armchair Psychology

Recall that in the first step in the Canberra two-step we collect the platitudes concerning the subject matter of interest, in this case beliefs and desires. These are the claims about beliefs and desires that reflect our ordinary usage of those terms. We then use these platitudes to define the theoretical role of beliefs and desires (Nolan 2009: 268). Is this actually what self-described moral functionalists do? Much hinges on the nature of these platitudes, and perhaps not enough has been said about them by the functionalists to generalize.

Smith (1994) says that platitudes are a priori truths: that two possible worlds alike in all of their natural features must be alike in their moral features is a platitude. Everyone agrees that the moral features of things do not “float free” of their natural features (21-22). A bit later he gives some examples of color platitudes: that there is no such thing as transparent matte white is a platitude, known a priori by competent users (29). Smith’s idea is that we come to master color vocabulary by coming to treat such remarks as platitudeous.

The issue I have so far been raising is one about the nature of such platitudes. What Smith and other Canberra Planners need is that platitudes concerning mental states like belief and desire be of the same kind as these exemplary platitudes (such as those about whiteness). I have argued that this is not the case. What is required for successful folk practice may be described philosophically by these platitudes, but it isn’t obvious that the folk must have mastery of these platitudes in order be largely successful at predicting and explaining behavior. I have said that my objection is roughly a version of Schiffer’s (1987), which Jackson and Pettit have
tried to address. According to this objection it might be true that philosophers consider the claim, “there is no such thing as transparent matte white” a platitude in the sense that speakers who violate might lack mastery of the color terms from the standpoint of a philosophical characterization of folk discourse (the philosophical perspective), but it need not be the case that such a person will fail to employ the concept with predictive and explanatory success in many cases (the folk perspective). These are two distinct issues.

So the crucial question now is whether philosophers really know a priori that folk knowledge (of at least the implicit sort, though perhaps also explicit) of functional roles is platitudeous. Remember that the functional roles as JPS describe them are the nature of the platitudes. It looks to me that their methods are a tacit concession that philosophers have no such knowledge a priori.

Let’s start with Smith. One of the crucial parts of Smith’s project is to defend a dispositional account of desire according to which some action is desirable under some set of circumstances if and only if we would desire to take that action in those circumstances were we fully rational (1994: 113, 177; 1997). Part of Smith’s argument for the view involves criticizing its main competitors, one of which is the phenomenological conception of desires. The phenomenological conception of desires basically says that desires like sensations are simply and essentially states that have a certain phenomenological content (105). Weaker versions of the idea hold that desires are like sensations in that they have phenomenological content essentially but differ in that they have propositional content too (108). Set aside the plausibility of the varying accounts of desire here and focus strictly on the methods that Smith employs to test them. Here is Smith:

Do we really believe that desires are states that have phenomenological qualities essentially? That is, do we believe that if there is nothing that it is like to have a desire, at
a time, then it is not being had at that time? I should say that, at least as far as commonsense opinion goes – and what else do we have to go on in formulating a philosophical conception of folk psychological states – we evidently have no such belief. Consider for instance what we should ordinarily think of as a long-term desire: say, a father’s desire that his children do well. A father may actually feel the prick of this desire from time to time, in moments of reflection on their vulnerability, say. But such occasions are not the norm. Yet we certainly wouldn’t ordinarily think that he loses this desire during those periods when he lacks such feelings (1994: 108-9; italics added for emphasis).

I have added italics to bring out what I take to be the most telling parts of the passage: references to behavioral norms. The argument looks very much like an appeal to systematic empirical observation precisely because it relies very heavily on what is normally the case.

We have to be careful here because there are two things going on in this passage at the same time: there is Smith’s reflection on what we evidently believe about the nature of desire and then there is his reflection on why we believe what we believe about desire. At first, it is just the latter that seems to require the empirical evidence. To know what we normally believe we have to reflect on what most of us think about the nature of desire by thinking through a hypothetical case of characterizing it in others (in this case in fathers). In fact though the former requires empirical consideration too, which is why Smith says that “evidently” we have no such belief about the nature of desire. What is evident are the norms governing the use of the term in the general population, or the way that people actually use the term when they use it in the same way that most everyone else does.

The first step in the Canberra Plan is to lay out the platitudes concerning the concept of interest, in this case desire. For Smith this amounts to saying what is crucial to something’s being a desire (1994: 31; Nolan 2009: 274n). So the relevant part of the above passage from Smith is the question whether we really believe that desires have phenomenological qualities essentially. And Smith’s answer is that we evidently do not really believe this because of the
way that most of us employ the term. That is, because of the norms governing the usage of the term “desire” in the linguistic community. But the only way that we can have access to those norms is to go out into the world and observe the way that people use the term. One cannot have access to linguistic norms, by which is meant here standards determined by linguistic practice, from the armchair. At most you can determine how people might use the term incoherently, as by using “bachelor” to refer to married men, for example. But that is not the same as determining how most people in fact use the term.

Of course we can respond on Smith’s behalf here that this is his criticism of an alternative conception of desire rather an argument for his own view. The argument appeals to linguistic norms and empirical evidence in an attempt to show that people do not really employ the phenomenological conception of desire, but this is not intended to be the first of step in the Canberra method, which is to round up platitudes.

The reply isn’t very helpful because in either case we are still just rounding up platitudes. Which version of desire best accounts for those platitudes is a question to be taken up later. As Smith uses the term “platitude” it should be an entirely a priori and armchair matter what is essential to the concept “desire.” In fact Nolan characterizes Smith’s view of platitudes as being “close to the nonnegotiable analytic truths end” of the spectrum of platitude characterizations (274). So the relevant consideration will be that people must use the term in such and such a way because to do otherwise would be incoherent, like trying to use the concept “white” by treating it as including transparent opaque matte white in its extension, or using “bachelor” to refer to married men. This isn’t the normal usage of “white” or “bachelor” on Smith’s view (I take it) precisely because you can’t use it that way coherently, not because as a matter of empirical fact people don’t. The former we can know a priori because it is an analytic truth, but
the latter requires some investigation. People ought not speak incoherently; *irregardless*, they very often do.

Second, the reply fails to help Smith anyway because his argument for the dispositional account proceeds by demonstrating a conjunction: the phenomenological account of desire is committed to problematic claim *and* the dispositional account is not committed to those claims. The second conjunct requires little argument: a dispositional account of desire is not wedded to the problematic claim (in particular, the epistemology of sensation) in virtue of the fact that it is not a phenomenological conception. So it is the first conjunct that shoulders the load. The argument against the phenomenological conception is most of the argument for the dispositional conception because it is the latter that Smith thinks leads the anti-Humean to the confused idea that an agent’s having certain beliefs may constitute her having a motivating reason (125). In other words, the platitudes are the platitudes, and the Canberra Planner always begins by rounding them up.

Smith needs to establish that philosophical psychology is in *need of* an alternative to the phenomenological conception of desire because of those platitudes. That alternative must be able to make sense of desires as states with propositional content, which in turn allows us to make sense of *commonsense* desire attributions (111). Smith thinks that there are two serious problems for the phenomenological conception, though here I focus on his argument for the first, which is that it gives an implausible epistemology of desire (108).

The idea behind the first objection is that any adequate account of desire must recommend an epistemology that is both prima facie plausible and that allows subjects to be fallible about their desires. He thinks that the phenomenological conception fails to meet both requirements. The epistemology of the phenomenological conception is based on the
epistemology of sensation, which holds that a subject is in a state with phenomenological content if and only if she believes she is in a state with that content. For example, she is in pain if and only if she believes that she is in pain. Consequently, “if we think of desires on the model of sensations, it is plausible to hold that a subject desires to $\Phi$ if and only if she believes that she desires to $\Phi$” (105). We can represent the biconditional as PD (which stands for “phenomenological conception of desire”):

\[(PD). \text{ A subject desires to } \Phi \iff \text{ she believes that she desires to } \Phi.\]

He argues that PD fails in both directions. From left to right it is plainly false, while from right to left it violates the fallibility constraint, i.e., any adequate conception of desire must be able to account for the fact that people are sometimes wrong about the desires that they have.

The interesting point for my purposes has nothing at all to do with whether PD is a plausible account of desire but only with the methods that Smith uses to upend it. Both of his arguments against PD proceed by counterexample.

He argues first that PD is false left to right because there are perfectly realistic cases in which it fails to hold. We are asked to consider John, a narcissist who each day buys a newspaper from one particular newsstand while en route to work. Though John might just as easily get the paper at any newsstand, he continues to patronize just the one stand in particular. As it turns out, that particular stand has mirrors placed in such a way that shoppers cannot help but look at themselves as they make their purchases. If we were to tell John that perhaps he buys his paper there because he enjoys looking at his reflection, he would vehemently deny it, “and it wouldn’t seem to John as if he were concealing anything in doing so” (106). Suppose however that without the mirrors, John would no longer visit the shop. From this Smith concludes that we have reason to reject PD left to right:
If all this were the case, wouldn’t it be plausible to suppose that John in fact desires to buy his newspaper at a stand where he can look at his own reflection; that, perhaps, he has a narcissistic tendency and buying his newspaper at that stand enables him to indulge it on the way to work? And wouldn’t it also be plausible to suppose that he does not believe that this is so, given his, from his point of view, sincere denials? If this is agreed, then we have reason to reject the principle left to right… (106)

I have no interest in defending PD here, so let us just concede to Smith for the moment that this case is rather plausible. The real issue anyway is not whether the case is plausible but whether the truths about the nature of desires that it reveals are platitudes about the concept that we find in folk discourse or practice. This is what Smith means when he says that we need an alternative account of desire in order to “make sense of desires as states with propositional contents and that thus allows us to make sense of our commonsense desire attributions” (111). Is Smith’s attribution of desire in John’s case really commonsensical or platitudinous?

I think it is not. We have seen that even for Smith platitudes are instances of knowledge, implicit or explicit, which are such that in coming to treat that knowledge as platitudinous we come to have mastery of the concept (31). Recall that Nolan characterized Smith’s view of platitudes as coming close to the nonnegotiable analytic truths end of the spectrum. It is surely not anything like an analytic truth that desires are such that they make us do things that we might not consciously realize we desire to do. If this is a truth about the nature of desire at all it is surely known a posteriori in virtue of some experience with desiring and watching other people desire.

But is it platitudinous anyway? I doubt it particularly because the example is so controversial. The mere fact that most would say to John that he frequents that particular newsstand because he likes to view his own reflection is evidence that that is what most people would think. To those of us not trained in psychiatry it would probably be very surprising if it were really the case that for at least the first few trips to that stand John had no idea why he
preferred it. And it would be even more surprising that John could lack any awareness whatsoever of his enjoyment in looking at his reflection. Even within psychology this strikes me as a significant claim about the nature of desire in need of some empirical support. That is probably why Smith argued by counterexample rather than just by asserting the obvious fact that desires are such that they make us do things even when we have no idea that we desire to do them. It is not an analytic truth. It is not a platitude. It is not even really a piece of common sense. The entire argument looks to me like it rests on some shoddy armchair psychology.

Smith rehearses a similar story, this time about John the aspiring musician, to undermine PD from right to left. The story goes like this. John says that one of his fundamental desires is to be a great musician. But his mother has put this idea in his head by beating into him a deep appreciation for the value of music. John willingly admits that he has a great desire not to upset his mother, though he denies that this fact explains his musical pursuits. John’s mother dies, and suddenly his interest in music vanishes and he quits his musical pursuits. Smith then poses the crucial question: “wouldn’t it be plausible to suppose that John was just mistaken about what he originally wanted to do and that, despite the fact that he believed that achieving excellence in music was a fundamental desire of his, it never was? If so, then we have reason to believe that the principle is false right to left as well” (107).

The same point stands. It is perhaps plausible (though I remain unconvinced), but the relevant issue is whether it is platitudinous of the concept. If not, then we have no reason to build the characteristic into the concept during the first stage, which is the conceptual analysis part of the Canberra Plan.

Let me be clear a about what I am not arguing here. I am not arguing that it is no part of the nature of desire that desires make us do things of which we are utterly unaware. This could
turn out to be true of desire and it will still be the case that we arrived at the truth via observation and assessment of cases. The one thing that seems clear is that someone (a member of the folk) could be perfectly adept at applying the term without any knowledge whatsoever of the fact that sometimes desires work subconsciously. Such a person would be exactly the type to criticize John for his narcissism in the first example. There is just nothing incoherent about a conception of desire lacking the subconscious element, and so it cannot be a platitude the accommodation of which is required for an adequate conception.

Second, and more importantly, I am not arguing that Smith will fail to abide by his own methodological standard any time that he appeals to a behavioral example or a hypothetical case. We can agree that a posteriori considerations have some role in conceptual analysis. It is just that any such considerations will have to reveal a rather obvious truth about the nature of the concept in question in order for it to legitimately reveal a platitude. And on the Planners’ account this will generally happen by considering what must be the case about the nature of desire in order for folk practice to be so successful.

But there is just nothing obvious about the purported features of desire that Smith takes these hypothetical cases to show, neither as an analytic matter nor as a fact about what the folk must do tacitly. He seems to concede this much when he asks whether the explanation for John’s behavior is plausible rather than declaring that it is obvious or necessary. And that is why I think Smith’s methods betray a need for empirical investigation. The whole point of engaging in empirical science at all is to test a posteriori claims that aren’t perfectly obvious. Successfully supporting a scientific hypothesis amounts to showing that it holds generally and so can be applied to help with other untested cases, precisely like John’s.
§6. More Shoddy Psychology as Conceptual Analysis

The problems so far presented are directed at Canberra Planners. I have focused just on Smith. It seems to me that his method violates his Canberra commitments. In fact, it has to, for otherwise he has no way to argue in favor of a dispositional conception of desire as opposed to a phenomenological conception. The considerations that reveal the shortcomings of the phenomenological conception are relatively controversial empirical considerations about why people do what they do. And this is a far cry from rounding up platitudes and showing that they all point to a single (dispositional) account of desire.

This criticism is thus not applicable to others who engage in the dispute about the nature of desire and moral motivation, at least not directly, because they don’t subscribe to the Canberra Plan. But because so many philosophers are like Smith in taking an almost exclusive interest in beliefs and desires as the centerpieces of moral psychology, I think a variant of the same problem applies. And it helps to underscore the importance of moving toward methodological naturalism.

Whatever their deeper commitments in the philosophy of mind or metaphysics, philosophers like McDowell, Platts, Little, and Nagel who engage in disputes about moral motivation argue about the virtues of various conceptions of beliefs and desires on the basis of similarly controversial empirical considerations. For the empirically minded among them this is no real problem because they will gladly concede that a posteriori empirical considerations are crucial to determining the best account of desire, motivation, etc. The problem is that this overlooks the idea that a genuinely empirical approach to understanding moral agency will likely result not just in more theories of the nature of belief, desire, motivation, and so on, but potentially in different and more complex accounts of mental states and processes in need of
explaining. In other words, they may be wrong to play the game as it has been laid out by an 18th Century theory of human psychology.

As an illustration of what I mean, consider Thomas Nagel’s (1970) case against Humeanism, the chief alternative to his own positive theory of practical reasoning. In one of Nagel’s widely discussed arguments, he objects to Humeanism on the grounds that it cannot account for prudential motivation (pp. 39-40). This, basically, is one’s having reasons in the present to prepare for fulfilling one’s future desires later on. We have prudential motivation when our current recognition that we will have a future desire motivates us presently to take steps toward fulfilling that future desire.

Humeans are committed to the view that an agent’s having a motivating reason to take some action requires a current desire to take that action. Consequently, Humeans can explain prudential motivation only by appealing to a present desire: an agent’s recognition that she will have some future desire must give rise to a present desire to take steps toward satisfying that future desire. So the standard Humean line—really, it seems the only line available for the Humean—is to claim that in cases of prudential motivation agents have a present general desire to further their future interests (Nagel 1970; Smith 1994).

Nagel objects to this explanation that on the one hand “it does not allow the expectation of a future reason to provide by itself any reason for present action” and on the other hand that “it does allow the present desire for a future object to provide by itself a reason for present action in pursuit of that object” (39). Thus it carries two unacceptable consequences: first that “I may have a reason now to prepare to do what I know I will have no reason to do when the time comes” and second that “I may have no reason to prepare for what I know I shall have reason to do tomorrow” (39-40).
In defense of the first, he offers his now widely discussed persimmon example. Suppose that on Saturday I have a desire to eat a persimmon on Sunday. That is, on Saturday I have a present desire (viz., a present desire about what to do on Sunday) to eat a persimmon, a future object. However, I am aware that come Sunday I will no longer desire to eat the persimmon. According to Humeanism I have in this case reason to buy the persimmon on Saturday, namely, that doing so will satisfy my present desire to further my future interests.

Intuitively this is problematic: buying a persimmon on Saturday for Sunday when I know that I won’t desire it Sunday is pointless and wasteful. Nagel’s data here is some combination of the introspective and the observational. It is an appeal for us to reflect upon our own psychological states, and perhaps our observations of others’ states, to recall analogous situations in which we, or others, have had desires for the future and simultaneous expectations about changes in those desires. It is fundamentally concerned with previously accumulated behavioral evidence. Whether such an example is conceptually possible depends on whether we can recall, or imagine based upon real experience, finding ourselves or others—real moral agents—in similar scenarios, having all the relevant psychological states. This is why we run headlong into conflicting intuitions when dealing with arguments that hinge on conceptual possibilities. Again, I am not arguing that there is anything wrong with this method in principle. The problem is only that if we are going to make substantive empirical claims about the cognition and behavior of real moral agents then it is a mystery why we should be content to confine ourselves in advance to the limited framework of belief-desire psychology just so that we can do psychology from the armchair.

Strangely, even when philosophers have seemed to acknowledge the controversial nature of the empirical claims they have ignored its importance. In a reply to Nagel’s persimmon case
given in defense of Humeanism, Maslen (2002) begins by pointing out just how psychologically unrealistic Nagel’s counterexample is only to insist that the consideration will play no part in his argument, and further that an unrealistic counterexample is a counterexample nonetheless:

Although it will not form any part of my argument against Nagel, it is worth noting that the persimmon example is contrived. It is difficult to tell a realistic and coherent story that explains both why Nagel had an unmotivated desire on Saturday to eat the persimmon on Sunday and how he knew in advance that this desire would cease on Sunday. A contrived counterexample is still a counterexample. But are there analogous realistic cases of this behavior (2002: 42)?

I disagree: when it comes to assessing arguments based upon empirical observation, a contrived counterexample is no counterexample at all. When I assert the existence of albino ravens as a counterexample to the claim that all ravens are black, the only relevant consideration is whether there really are albino ravens. If my claim is contrived in this context it is utterly worthless.

It seems that the only reason that Maslen, or anyone else, would say that contrived and unrealistic counterexamples are nevertheless counterexamples in this context is if they were working under the assumption that empirical considerations properly have no bearing on folk psychological concepts. Otherwise we should expect the outcome of the dispute to depend entirely on the instantiation of such counterexamples in real moral agents. So it seems Maslen’s disclaimer is actually rather revealing. If the realism of the example plays no part in the argument, why raise the matter at all? If, as Smith says, all we have to go on in formulating a philosophical conception of folk psychological states is commonsense opinion, then whether Nagel’s counterexample is utterly contrived is irrelevant, and indeed Nagel needs no example at all for his point to stand. Why, then, did he include it, and why is has so much been said toward its assessment? For that matter, why have any of these philosophers—along with countless others I haven’t discussed here—bothered to appeal to these putatively hypothetical behavioral considerations at all?
Platts (1979) makes a similar move. He argues that the Humean theory of motivation is committed to the view that we must introspect the presence of desire whenever we are motivated, and that as a matter of empirical fact this is evidently false: “At first sight, it seems a painful feature of the moral life that this premiss is false. We perform many intentional actions in that life that we apparently do not desire to perform. The difficulty of much moral life then emerges as a consequence of the apparent fact that desiring something and thinking it desirable are both distinct and independent” (256; quoted in Smith 1994: 110). There is a good deal of talk about observation, appearances of real life, and facts of the matter here.

§7. Some Quasi-Naturalistic Approaches

A related objection applies to many self-described naturalists, who sometimes implicitly endorse the empirical assumptions at work in traditional philosophical disputes. And many of them never quite make it out of the armchair.

Jeanette Kennett (2002; Kennett & Fine 2009) has been one of the first naturalists to try to invoke substantive empirical considerations explicitly for the purpose of revising our conceptions of moral agency. Ultimately, though, her project loses touch with the purpose of naturalism by inadequately employing controversial empirical data in defense of an old philosophical theory.

---

37 The extent to which Kennett’s project is naturalistic is difficult to pin down, in part because her conclusions have changed considerably. Her 2002 paper contains indications that she intends to restrict the discussion to basic empirical observations about psychopaths and autistic people and thus to something like commonsense folk psychology. Even so, it contains a substantial review of empirical work, even if that work relies on dubious methods (which, at points, she admits). Her more recent arguments (e.g., Kennett & Fine 2008), however, indicate that she has moved entirely toward the naturalistic project of employing systematic empirical data. But even there she seems committed to the likelihood of vindicating traditional philosophical theories.
Kennett (2002) offers a lengthy argument in favor of a Kantian account of moral motivation, according to which moral judgments are necessarily motivating, over a Humean account using observational data from psychopaths—individuals suffering from antisocial personality disorder (APD)—and autistic patients. Her argument hinges on the claim that while both psychopaths and autistic patients lack empathy, an emotional or conative state presumably essential to moral life, only the psychopath and not the autistic person lacks the ability to act on her considered moral judgments. In defense of the claim she appeals to some psychological literature, most of which involves observation (e.g., from well-known physicians such as Oliver Sacks) and self-report from well-known autistic individuals (e.g., Temple Grandin) to show that indeed both psychopaths and autistic persons lack empathy, and further that most autistic persons and not psychopaths are nevertheless capable of moral motivation.

Here is one such passage:

However, high-functioning autistic adults and those with Asperger’s syndrome can achieve moral autonomy by other means. It appears that they can develop or discover moral rules and principles of conduct for themselves by reasoning, as they would in other matters, on the basis of patient explicit enquiry, reliance on testimony and inference from past situations. Temple Grandin describes having a library of experiences in her head, ‘videos’, which she would replay again and again to learn how people behave (2002: 351-2).

If she is right, then it seems we have evidence in favor of one particular account of moral motivation inside a folk psychological framework. Moral judgments, it appears, are capable of motivating even in the absence of conative states, in this case empathy. Kennett thinks that the data bear this out. What really matters for settling the dispute, then, are not the constraints set in place by the folk definitions of putative psychological states but the data concerning the

---

38 Kennett claims that “anti-social personality disorder” has replaced “psychopathy” in the psychiatric literature. Here I am just describing her view. There may be reason to contest the identification of these psychological conditions (see Hare 1998 for example), but I take no position on that issue here.
instantiation of these putative states in real moral agents. This naturalistic approach seems more promising, but what is so surprising is that she carries the results directly back to a commonsensical, folk framework, without offering any independent reason to think one or the other theory in that framework must be correct. In fact she seems so concerned to preserve the traditional framework and defend the Kantian theory that she sacrifices methodological rigor to do it.

Her goal is to use behavioral observations in defense of a traditional philosophical thesis. But if behavioral observations are to be used in the formulation of psychological generalizations, they must be made within the framework of standard experimental protocols with the proper controls in place, they must be repeatable, and the data collected must have enough statistical power to permit generalization to larger populations. Much like casual reflection on one’s appetite for persimmons and conceivable attitudes of fathers toward their children, the introspective reports of Temple Grandin (whose reflections, incidentally, concern a disorder that has now been removed from the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders [DSM], namely Asperger’s syndrome\(^{39}\)) fail to satisfy these criteria. Genuine methodological naturalism should start with a commitment to building theories around well-supported empirical hypotheses. For that reason, selectively attending to reflections from notable scientific figures to vindicate traditional philosophical theories should not be regarded as a genuine form of naturalism.

More recently, though, Kennett & Fine seem to have revised their position to accord better with naturalistic methods. Here is what they now say about the relationship between the empirical science and the traditional philosophical theories:

In recent years there has been a surge of research into moral cognition and judgment in the social and cognitive sciences and the hope has been that this work might inform philosophical debates… Do [dual processing models of cognition]—can they—offer empirical vindication of one or the other philosophical position (2009: 78)?

Here Kennett and Fine seem more open to the idea that traditional theories may be entirely wrongheaded, which it seems to me is a step in the right direction for naturalism.

§8. Toward the Science of Moral Motivation

I take these last few sections to establish the importance of systematic empirical observations to the project of developing an accurate, empirically respectable theory of moral motivation. It’s not just that moral philosophers rely on empirical considerations but that without them they have nowhere to go. In recognizing this tacit commitment to empirical inquiry we should start to see that there is something deeply confused about making (or claiming to make, in the case of the Canberra Planners) philosophy in some significant sense prior to science when it comes to arguing about the nature of moral agency. And that is the real motivation for defending methodological naturalism. The physiological constraints that the sciences are now placing on theories of human cognition and behavior are not going to disappear because we ignore them. Better to confront them first.

The question that moral functionalism poses is whether this naturalistic move is really necessary anyway. If the Planners are right, then what I have failed to grasp is the philosophical conclusion that we must be able to arrive at truths about the nature of moral agency using a
functionalist theory of the nature of mental states. That is, using an account of mental states in terms of the *folk roles* established by the myriad platitudes that describe folk practice.

To finally make good on my promise, then, I turn to the empirical sciences. The thesis I defend in the remaining chapters is this: if it is really beyond dispute that the future neurosciences *must* invoke the folk roles in the relevant explanations of cognition and behavior for the reasons the Planners give, then we should certainly expect that a look at the *current* neurosciences will do little or nothing to threaten the importance of those roles in the relevant research paradigms.
4. Operational Definitions in Philosophical Psychology

§1. Preliminaries

In this chapter I begin some preliminary work on the question whether the folk roles for belief and desire have what I have been calling “theoretical realizers,” and hence whether they might also have physical realizers, in the empirical science of moral judgment and motivation. This paves the way for the second horn of the proposed dilemma for the Humeanism dispute, which challenges directly the moral functionalist’s claim that “future neuroscience” must ultimately preserve intentional mental state concepts.

In the last chapter I argued that what the functionalists need in order to vindicate the claim is evidence not just to the effect that the folk roles of beliefs and desires play some role in neuroscience in general but rather that those roles must serve in the relevant capacity in the neuroscience decision and motivation, and also the very recent cognitive neuroscience of moral judgment, which are the developing branches of neuroscience most relevant to the philosophical dispute about moral motivation. In order to determine whether this is a likely outcome it will be necessary to engage with those sciences directly. The ultimate goal is to determine how philosophical explanations in moral psychology compare to explanations in neuroscience.

At present, however, the most glaring obstacle to achieving that goal is the absence of clear and consistent theoretical terms in the philosophical literature. Philosophers have systematically failed to define key terms in their arguments for their theories of moral motivation. Here I briefly consider both the problems that arise from that failure and how we might begin to solve them.
§2. The Absence of Operational Definitions in Moral Psychology

There are several difficulties with trying to determine the veracity of a philosophical-cum-empirical theory of moral agency by looking at the data and methods of current neuroscience research. Foremost among them is that the philosophers generating the theories consistently fail to tell us what to look for. It is considerably easier to make vague remarks about “future science” than to give some idea about how or where a mental state construct might actually be involved in the theories developed by the neural and cognitive sciences. The vagueness carries the advantage of making the claim virtually impossible to refute. For any plausible piece of evidence that a theoretical construct employed in an armchair theory (e.g., “moral belief”) is absent, trivially true, wrongly located, etc., in the science, the philosopher can reply either by emphasizing their interest in future neuroscience or by simply insisting that that (whatever construct pointed to in the empirical project) wasn’t what was meant by the construct in the armchair theory.

Philosophers who talk loosely about the future of neuroscience give us absolutely no independent reason to think that the future will be significantly different from the present. And in the absence of that evidence, we should expect that the future would be different from the present only in degree and not in kind. Future neuroscience isn’t under construction in a vacuum somewhere. Rather, it will be built upon the foundations that are already in place. If the philosopher’s theoretical mental state constructs are as fundamental as those like Jackson, Pettit, and Smith claim they are then there should already be some evidence of their presence. So we should take a careful look at present neuroscience first. One of the main reasons it has been so difficult to do that is that there is a complete absence of what we might call, to adopt a controversial term, operational definitions. In this chapter I will use the term in a weaker sense
than is meant in scientific literature, but it will be helpful first to say something first about the standard scientific usage.

Operational definitions are simply definitions of a phenomenon of interest that unlike commonsense terms make possible objective, systematic and consistent observation. They are crucial to systematic empirical inquiry because they help to ensure some degree of consistency across repeated measurements of variables, both within and across experiments or observations. For example, psychologists interested in the relationship between intelligence and year-round schooling must recognize that while these concepts seem clear enough for everyday purposes—most of us have some idea about what they mean and can probably form some ideas about the possible relationships between them—they are unsuited for generating consistency among observers. There is a serious risk that different people, even professional psychologists, mean different things by the terms. Some might take intelligence to be equivalent to a score on the Stanford-Binet test while others take it to be equivalent to a score on the Wechsler Adult Intelligence Scale (WAIS). Consequently the validity of the results will depend on the extent to which those tests are interchangeable instruments — a controversial empirical question in itself.

To avoid this problem in empirical research scientists use precise operational definitions, which specify the methods that were used to produce a phenomenon or to measure a variable. Interest in intelligence and year-round schooling becomes interest in the relationship between scores on, say, the Stanford-Binet and a 12-month school year. Despite sharing with empirical scientists the goal of establishing legitimate empirical generalizations, Humeans and anti-Humeans have done almost nothing to standardize their terms. However obvious the point seems, it has been almost entirely ignored in the philosophical literature. In this chapter I want to address this problem.
To the extent that Humeans and anti-Humeans plainly appeal to empirical observations to defend their chosen theories, operational definitions in the above sense of the term will seem relevant enough: we need to be clear and consistent about how observations are to be interpreted, and that requires some standardization, or operational definitions. Yet some will object that insisting on the need for operational definitions in this sense goes wrong in at least one of two ways.

First, one might deny that Humeans and anti-Humeans are even in the business of defending their theories on the basis of empirical observation rather than, say, conceptual analysis, intuition, or still some other kind of philosophical data prior to empirical investigation. In this case, operational definitions in the scientific sense outlined above will be irrelevant. Philosophy, the objection continues, is prior to science in the sense that disputes about proper “operational definitions” are themselves philosophical disputes.

Second, one might object that many of the participants in the Humeanism dispute, including Jackson, Pettit and Smith, take a functional approach to the concepts that I claim stand in need of operationalizing. In giving a functional characterization of these concepts, such as ‘belief’ and ‘desire’ and ‘motivation,’ functionalists have provided not only a sufficient alternative to operational definitions but in fact the very alternative that emerged from the demise of analytical behaviorism.

In part I addressed the first objection in Chapter 3. There I argued that proponents of Canberra-style functionalism sometimes disguise appeals to empirical observation as the deliverances of armchair psychology. But we can set that argument aside for the moment since by operational definitions here I mean to refer to something weaker than the scientific conception. The idea that I will defend in the rest of this chapter is that participants in
philosophical disputes like the Humeanism dispute run the risk of talking past one another in the absence of consistent terminology. Whatever the type of data invoked, empirical or philosophical, there is not much use in arguing about the status of theories of the causal efficacy of beliefs prior to determining that we mean the same thing by the concept ‘belief,’ for it may be nothing more than a definitional dispute that drives our disagreement about causal efficacy. This is, as we will see shortly, as much as my request for “operational definitions” requires. I mean quite simply that we must have consistent definitions with which to operate, in particular, when arguing about what the definiendum (or definienda) does.

As for the second objection, according to which functional definitions are sufficient to supplant operational definitions, the objection has more teeth in theory than in practice. In principle it seems right to point out that functional definitions provide all that we could require of operational definitions. But, at least in the case of functionalists like JPS, it is rarely the case that the functions of all of the concepts crucial to a particular dispute, such as the Humeanism dispute, are ever fully spelled out despite the fact that those functions are not simply obvious to the rest of us. It is true, as we have seen, that JPS are clear that they understand some cognitive states such as beliefs and desires in terms of their “folk roles,” the functional roles they play in folk theory. Even so it remains unclear how the appeal to folk roles automatically yields a story about the functions of such states that is complete enough to help us avoid confusion about exactly which states in some organism should count as beliefs or desires in some context. And, perhaps more importantly, note that nowhere do they answer crucial questions about the other relevant concepts in a given dispute. This problem is most glaring I think in the cases of the concepts ‘motivation’ and ‘moral judgment.’ For example, how are we to understand ‘motivation’ functionally? What is its folk role? This is not to deny that JPS will be able to
provide an adequate definition in functional terms but only that that answer is far from obvious. Of course, once that functional definition has been fully spelled out, we will have something near enough to an operational definition, and the argument in this chapter will have been given an adequate reply. Until then, we should worry about consistency in the Humeanism literature for the reasons to follow.

Humeanism is a theory of moral motivation. It holds that moral judgments are insufficient for motivation, since motivation requires in addition to the judgment the presence of a desire or similar conative state. Anti-Humeanism holds that the moral judgments themselves are sufficient for motivation either because they are somehow intrinsically motivating or because they necessarily trigger desire or desire-like states whenever they occur. Both are theories about the causal efficacy of a particular class of mental states, namely moral judgments. Anti-Humeans ascribe to moral judgments a particular causal property—the ability to generate motivational states either directly or indirectly—that Humeans do not.

For this reason both Humeanism and anti-Humeanism are something like attempts at formulating law-like generalizations about motivation in moral psychology, often on the basis of observational experience and thoroughly empirical considerations. Some philosophers who defend a view in this debate have thus far dodged the charge of engaging in shoddy science by restricting their claims to the domain of commonsense FP or to moral functionalism, which has commonsense FP at its core. In both cases the idea seems to be that because folk discourse is commonsensical we need not worry about the sciences undermining it. Intentional mental state concepts will undoubtedly be preserved by future neuroscience precisely because they are so commonsensical.
I have thus far suggested, in chapters 2 and 3 respectively, that each of these moves is problematic. Commonsense FP involves a kind of behavioral ascription, explanation and prediction—a folk practice—that may be useful for everyday purposes, but we don’t understand it well enough (at least not yet) to simply adopt its terms and concepts into an empirical theory of moral agency or to know where to find those concepts in the sciences.

Set aside for the moment that Canberra functionalists sometimes appear to violate their own methodological commitments by smuggling empirical generalizations into putative platitudes analysis. Whether we arrive at the need for postulating the folk roles of belief and desire in moral motivation by rounding up and analyzing platitudes a priori or by a posteriori empirical observation, the concepts that have emerged and which now occupy center stage in the moral motivation literature are unsuited to systematic and objective empirical observation and measurement. Consequently we should certainly not expect to find these concepts in the sciences. If we do, it will tell us much more about the current state of scientific practice than about the reach of folk wisdom.

This explains why the claims about cognition and behavior put forward by moral psychologists (like Smith, Nagel, Platts, and Kennett, which I reviewed at the end of the last chapter) turn out to be controversial. They are plausible, but hardly platitudeuous. For example, we simply don’t know for certain whether people like John, the narcissist in Smith’s examples, will patronize a mirror-laden newsstand more than a similar stand without mirrors because we haven’t clearly isolated the behavior of interest and tested for it in the population at large. Moreover, if it turns out that as a matter of empirical fact people in general are like John in this respect, or alternatively if we are just interested in John’s peculiar behavior as perhaps Smith is, then we still won’t know exactly why John so prefers the mirror-laden newsstand, though we
might begin to test hypotheses by specifying a particular cognitive entity or pattern, defining it in a way that makes it publicly observable and measurable, controlling for alternative explanations, and then testing for its presence. Until then, we have done nothing but speculate about human psychology, or contrive examples as Maslen (2002) rightly observes, on the basis of how we tend to employ a few commonsense terms like “belief” and “desire.”

Humeans and anti-Humeans have not yet come to an agreement on what precisely is meant by terms like “moral judgment,” “moral motivation,” “belief,” “desire,” and so on, but remarkably they all seem to agree that the terms will nevertheless have to be preserved by “future neuroscience.” Remarkably, they have actually done something like the opposite of operationally defining key terms: they have developed rather fluid mental state concepts, which can be molded to fit the present evidence.

Smith (1994), for example, reviews two competing conceptions of desire before advocating the one that better elucidates his distinction between motivating reasons and normative reasons (pp.104-111). However incisive the philosophical analysis, the mere fact that he means something different by the concept than do his opponents rules out the possibility of objectively observing the role that desires play in generating motivation, which helps to explain why the appeals to hypothetical behavioral cases (like John’s) are controversial. But as we have already seen this hasn’t stopped functionalists or even naturalists from appealing to instances of real moral agency to help settle whether as a general matter moral judgments motivate only in the presence of desire.

---

40 Jackson and Pettit (1990) claim this explicitly, but anyone who engages in disputes about moral agency, such as Humeanism and internalism, must be committed to the idea implicitly because they are in dispute over which is the more accurate theory of moral motivation. Exceptions will include only those who defend the idea that neither is accurate and those who insist on the folk perspective as discussed in chapters 2 and 3.
Any attempt to generalize from empirical observations made without the use of operational definitions is just bad science; the use of unsystematic observational data is just controversial, anecdotal evidence. In either case the results of the investigation fail to generalize and so surely will not help to settle the question whether, in general, moral judgments are sufficient to motivate. Unless, of course, the question is not an empirical one at all, in which case the myriad appeals to the behavior of real moral agents must be dropped (or more to the point probably wouldn’t have been made in the first place). This is the conclusion drawn at the end of the previous chapter.

In the case of JPS’ functionalism, the dearth of operational definitions for the putatively requisite functional roles is a serious problem because it leaves us entirely unable to evaluate two of their central claims. First, we have no way to evaluate empirically the class of claims made by functionalists concerning the so-called folk platitudes (such as what they are like and how people actually rely on them) and other supposedly obvious truths about real human cognition and behavior (as in the claims about John just discussed). Second, we have no way to evaluate whether the theoretical roles supposedly played by beliefs and desires in the production of the various mechanisms of moral agency (such as an agent’s making a moral judgment or being morally motivated) really are explanatory from the standpoint of the sciences whose goal it is to explain some of the core elements of philosophically interesting phenomena, in this case the mechanisms of motivation or motivational behavior.

Canberra functionalists like JPS have insisted that this point is irrelevant. The folk roles, they might say, do not need to be shown suitable for empirical inquiry because they are the things around which that inquiry is built – in some sense they make that inquiry possible. Something has to make folk prediction and explanation successful, and whatever might turn out
to be the occupier, it is the something-role that has to appear in science. Finding out what might play the role is the second stage in the Canberra two-step.

My main line of argument against this point (in chapter 3) has been that we know too little about the way in which the role is played to assume that the role referred to by the folk concepts ‘belief’ and ‘desire’ is in fact required by folk practice, which is in turn supposedly captured by the explanations we get from empirical science. It is not enough to know only that some role must be played to bring about the phenomenon of moral motivation. A non-trivial theory of moral psychology will have to tell us something about how the role is played, and only then will we be in a position to determine what might play it in the relevant sciences.

Here I want to make the point in slightly more technical terms: even functionalists need to provide operational definitions for the folk roles that they insist must appear in future scientific theories. Those definitions might take a slightly different form than a typical operational definition (e.g., intelligence as a test score) because they are concerned strictly with a functional role, but that doesn’t thereby eliminate the need for consistency or clarity in claims based upon observations of the way in which the role is played.

Before proceeding we should consider some of the more common objections to the use of operational definitions in general from various philosophical camps. One objection, adapted to the intelligence example, goes something like this: “WAIS scores might be correlated with success in school, but we’re interested in intelligence, not WAIS scores; when I use the concept I intend it to be something much richer, or at least broader.” From this it is concluded that the

---

41 The versions of this objection that come to mind are often about the operational definitions used in animal models of human behavior in neuroscience and psychology research. For example, memory in rats might be defined operationally as efficiency in running a particular maze. Critics observe that this fails to capture the depth of our concept of memory in everyday contexts. That may be true, but my point here is that criticisms of this operational definition of
procedures of empirical science are unhelpful for solving the philosophical question under discussion.

The objection is not well founded. It is admittedly easier to see how it goes astray in a toy example than in a case of genuine philosophical dispute, such as the Humeanism dispute, but the problem in both cases is the same. We are perfectly free to disagree about operational definitions, or to assert that better operational definitions are available for any given concept. This is certainly ground for a legitimate methodological dispute. Moreover, one is certainly free to criticize every operational definition proposed for any given variable. If the criticisms are good, then it may be that at present we have no good way to study the variable of interest, or we may decide that the research in question lacks scientific validity (i.e. fails to measure what it purports to) as a consequence of its employing bad operational definitions. But it doesn’t follow from this that we are therefore free to return to using vague folk terms or concepts (even if they are strictly functional) to settle empirical questions, nor does it follow that the norms of scientific practice are no longer required in order to make empirical generalizations.

Another objection says that to give an operational definition is to do substantive philosophical work prior to science. The process of operationally defining a term or concept is (involves, or is very much like) the process of conceptual analysis. This is perhaps why the Canberra Planners are inclined to preserve a crucial, prior role for philosophy: if we want to understand what we need to study, we have to do some conceptual analysis because that is precisely how you determine the nature of the concept or its role in everyday practice.
One who objects this way is right to note similarities between conceptual analysis in philosophy and operational definitions in science but wrong to claim that they are identical processes. Operational definitions in a scientific research paradigm are simply ways of determining whether two studies which appear to be about the same phenomenon, say episodic memory, are actually about the same phenomenon. In this sense they are a lot like stipulations.42 A research group specifies how they defined the concept of interest for observation and testing and the proceeds to explain the test and the results. Any claim made about the phenomenon on the basis of the results is a claim about the operational definition, not about the broader concept of interest.

Memory researchers put rats in a maze because they are interested in memory, but the claims that they make about memory are made about the concept as operationally defined, not about memory simpliciter or memory in some colloquial sense. This is usually made plain right in the titles of their papers. Here, for example, are a few of the top results returned for a search with the keyword “memory” in PubMed’s database (accessed July 19, 2010). I have italicized the operational definitions:

Mistry, R. et al. (2010). Dentate gyrus granule cell firing patterns can induce mossy fiber long-term potentiation in vitro.

Kozyrev, S.A. & Nikitin, V.P. (2010). Neuronal mechanisms of reconsolidation of an associative aversive skill to food in the common snail.


42 Though they are generally in accordance with the stipulations of researchers who study the same phenomena. The point here is that even if the stipulations aren’t in agreement, we can tell just by looking at the research methods and then quickly determine that we have two scientists interested in explaining different phenomena.
Notice that you could replace the italics in each case with the colloquial term ‘memory’ since that is ultimately what each of these studies is about. This is essentially what is happening in the moral motivation literature. Canberra conceptual analysis is supposed to be a way of arriving at substantive truths about the concept itself on the basis of rounding up platiitudes. That is not just a matter of stipulation. If platitude analysis reveals a truth about the concept bachelor, e.g., that a bachelor is an unmarried man, then presumably anyone who uses the term differently uses it incoherently and ought to be corrected. In a case like this it is not simply that two people are choosing to stipulate a different definition as in a disagreement about some philosophical definition like “supervenience.” By contrast, there is no dispute between Mistry et al. and Novitskaya et al. above, and neither is using the concept ‘memory’ incoherently. They are just studying two different things, which are ultimately related under the umbrella of memory, and are quite clear about that fact.

Of course, there are similarities between conceptual analysis in moral psychology and operationally defining terms in empirical science. The most important one is that intuition and common sense will play a major role in the operational definition that researchers choose. Crucially, though, once the commonsensical work is done and the operational definitions are in place, the experimental results apply to the operationally defined variables and not to the folk concepts. As a result, the objection—that because commonsense analysis is prior to the science, the science is therefore in its nature philosophical—is confused. Once operational definitions have been employed, they are closed to philosophical reinterpretation (though of course they are always open to philosophical challenge). The data and its implications may be open to interpretation, but the variables themselves are not. This is the sense in which providing operational definitions is not the same as doing analytic philosophy.
Strangely, even self-described naturalists often overlook this last point. Philosophers who seek to apply scientific data to traditional philosophical problems run afoul of this basic methodological point because they have little choice but to reinterpret existing operational definitions so as to make them relevant to the pre-established philosophical concepts (e.g., Kennett 2002; Kennett & Fine 2009). Or, worse still, experimental philosophers design their own experiments to investigate philosophical concepts without explaining (either to their research subjects or their readers) the operational definitions of the concepts of interest.

For example, Nichols (2002) collects intuitions on the moral judgments of psychopaths without explaining to his subjects the criteria for diagnosing psychopathy in the psychiatric community. Of course, that isn’t a problem if the goal of the study is just to round up opinions about the definition of ‘psychopath,’ but it is a problem when the point of the study is to determine the relationship between psychopathic judgment and motivation because everyone surveyed may have a completely different idea about what is meant by these terms. As a result, we have no way to know whether the intuitions we’re collecting are really about the same thing.

To claim, for example, that the science of motivation vindicates a Kantian theory of moral agency without offering a precise operational definition of the Kantian theory—that is, a version of the theory which employs concepts that to some reasonable degree match the variables of interest in the empirical research—is a lot like claiming that year-round schooling increases intelligence on the basis of a correlation between performance on the Stanford-Binet and children enrolled in schools with 12-month academic calendars – a passable headline in science journalism, probably, but a methodological problem in more technical work. Or like claiming that neuroscience has discovered the “God part” of the mammalian brain on the basis of some fMRI scans of subjects engaged in religious or spiritual behavior (e.g., Alper 1996). This is not
to deny that “spirituality” (whatever behavior that might involve) is a brain product, but simply to point out that reinterpreting the operational definitions employed in scientific research to make them fit the colloquial terms with which we began is a serious methodological mistake, one which sometimes makes empirical research appear more philosophically relevant than it is. Naturalists, I think, have become comfortable violating this point.

All of this ultimately raises a concern about question begging. One of the most serious problems with the moral motivation literature at present is that because of the methodological problem just discussed it seems the only way to vindicate one theory over other is to beg the question. Failure to offer operational definitions means we can simply define the concepts of interest in whatever way best accommodates the data.

For example, if Jones is an anti-Humean and Brown points out to Jones that as an empirical fact psychopaths judge that killing is wrong but do it anyway, Jones’ best response is simply to deny that the psychopaths have actually made genuine judgments. In turn, Brown can simply reply to Jones that the psychopaths must be making genuine judgments since they can accurately describe all of the wrong-making features of the act of murder.

Exactly this dispute has given rise to a discussion about the so-called “inverted commas” sense of moral judgment described by David Brink (1986) and Hare (1952). According to philosophers like Brink, “inverted commas” moral judgments, which take their name from the use of single quotation marks to indicate that a word or expression is intended in the way others and not oneself would use it, are insincere moral judgments that simply indicate familiarity with typical appeals to moral properties. That is, the claim that killing is wrong when uttered by a psychopath (who we are to assume enjoys killing, I take it, because we know how to wield the concept in our daily discourse) becomes: ‘killing is wrong.’ The inverted commas (or single
quotation marks) indicate that the psychopath is insincere to the extent that she is reporting her own feelings. Really, she is just reporting what is generally meant by the claim, or what others would say about killing. My point here is that arguments of this sort, which just beg the question, tell us much more about the state of naturalism and moral psychology than they do about psychopathy.

I am not criticizing legitimate philosophical disputes about the best way to define “judgment,” nor am I denying that such disputes make (or could make at any rate) important contributions to our understanding of moral agency. My point here is that any such dispute about basic definitions must come prior to the use of empirical data to defend a particular philosophical theory of moral agency such as Humeanism or anti-Humeanism. Once we have agreed on a standard use of the term, we can of course appeal to empirical considerations and defend our preferred theory of moral agency. We can of course also use empirical considerations to help determine how best to employ concepts like “judgment” in our theories.

But this is rather different from using empirical considerations to defend an internalist or externalist sense of the term, for example, which is just to argue for internalism or externalism about moral motivation. Insofar as we attempt to settle a dispute about the proper definition of “judgment” from within a dispute about which is the most empirically accurate theory of moral motivation, we are really just begging the question. This is at bottom the difference between the philosophical contribution to consistent empirical observation and the philosophical misuse of empirical data.

The following passages from the moral motivation literature, it looks to me, contain violations of this basic point. Many are taken from discussions of inverted commas judgments. Notice how in each case the crux of the issue is how properly to understand the concept
“judgment,” though in each case the argument is assumed to be about plain empirical facts concerning real human agents (like psychopaths) and the nature of their moral motivation:

There are, after all, real-life sociopaths like Robert Harris, the thrill-killer whose story is faithfully retold by Gary Watson (1987). Harris claims that he knew that what he was doing was wrong and that he simply chose to do it anyway; that he felt no conflict. It therefore seems quite wrong to suppose that he suffered from weakness of will, or, perhaps, from any other kind of practical irrationality either... amoralists do not really make moral judgments at all. Even if they do use the words to pick out the same properties that we pick out when we use moral words, they do not really judge acts to be right and wrong... (Smith 1994: 67).

The internalist must dismiss the amoralist challenge as incoherent. We may think that the amoralist challenge is coherent, but this can only be because we confuse moral senses of terms and “inverted commas” senses of those same terms. People can be unmoved by considerations which are only conventionally regarded as moral, but a genuine amoralist is inconceivable... The problem for internalism is that it does not take the amoralist’s challenge seriously enough (Brink 1986: 30).

In light of this, I carried out a preliminary study in which I presented philosophically unsophisticated undergraduates with questions about whether a given person really understands moral claims... These responses suggest that, at least in some populations, common conception of psychopaths is precisely that they really know the difference between right and wrong, but they don’t care about doing what’s right. Prima facie, this counts as evidence against the Conceptual Rationalist’s inverted-commas gambit. For it seems to be a platitude that psychopaths really make moral judgments (Nichols 2002: 289).

The point is not that amoralists really make judgements of some other kind: about what other people judge to be right and wrong, for example. The point is rather that the very best we can say about amoralists is that they try to make moral judgements but fail (Smith 1994: 68).

Motivational externalism contrasts with “motivational internalism” which is either the view that our moral judgements are partly constituted by moral motivation, or else that they would be if we were rational. The major problem for motivational internalism—in either guise—is that it flies in the face of common observation and first-personal experience of the fact that we can, without irrationality, be indifferent to morality... the phenomenon of indifference encourages motivational externalism (Zangwill 2003: 143).

Each of these, I think, just begs the question, so the overall lack of progress on the issue should be unsurprising. To understand how well the Humeanism dispute tracks the current empirical science, we have some basic questions to answer about our variables of interest, particularly the
concept ‘judgment.’ I want to briefly consider how we might begin to address these questions. Answers are long overdue.

§3. Operational Definitions for the Humeanism Dispute

The dispute is one about moral motivation. This is crucial because it will help us to get clear on what philosophers have meant by moral judgment, an ambiguous term that could be construed in a couple of different ways. On the one hand, a judgment might refer to the sort of state I’m in when I perceive a stimulus or situation and then react to it internally. This, for example, would be the kind of reaction involved in “passing judgment.” It might amount to labeling or categorizing a state of affairs, as when, for example, we evaluate (perhaps silently or internally) the quality of a food or even a philosophical argument. Call a judgment of this sort judgment1.

There seems to be an alternative sense of judgment that pertains much more directly to agency. For example, I might judge that giving aid to feed the starving is right to do, and in this case that judgment bears directly on a decision. I must decide whether to help or abstain from helping. In either case I make a decision, even if the decision is to do nothing or stop thinking about it. In short, these judgments are judgments that force conclusions about the kind of moral agent that we are – they concern what to do or abstain from doing. Call a judgment of this type judgment2.

It might seem that judgment1 is different from judgment2 in that it has no real bearing on motivation or action, at least not in any obvious or direct way. This is relevant because one way to block the claim that moral judgment is a species of decision would be to claim that the proper construal of moral judgment is judgment1. If we can coherently maintain that such personal,
internal judgments are the types of judgments relevant to moral agency then the idea that moral judgment is a species of decision loses plausibility, and the neuroscience of decision-making will be irrelevant.

I doubt the promise of this move. First, Humeans and anti-Humeans agree that their dispute is one about moral motivation. And motivation is always motivation to do something. For this reason it would be hard to make sense of the Humeanism dispute as a dispute about judgment. There is no sense in arguing about the source of our motivation to act morally if there are no available actions to be taken. And, second, it seems to me that most or all instances of judgment can be straightforwardly translated into instances of judgment without any substantive change of meaning anyway. Most any judgment is just an attempt to decide whether a particular noun is worthy of a particular adjective, as in the case of providing opinion, or to decide whether something is worthy of membership in a particular class. Judges of sport decide winners, judges in the legal system decide criminal responsibility, and judges of character decide how to describe or categorize others. However subtle, these are all instances of action, both verbal and nonverbal behavior, and from the standpoint of neurobiology they eventually get connected up with the recruitment of motor pathways.

So the kinds of judgment at issue in the Humeanism dispute must be those that bear directly on, or perhaps force, decisions about what it is best or optimal for the agent to do. Judgment within the context of the Humeanism dispute is best considered a species of decision - decision about what it is best to do.

We can make this idea more concrete by adopting language that links it directly to operational definitions already in use. Because “doing” implies various kinds of motor activity initiated toward various ends, we can safely say that the kind of decision relevant to the
Humeanism dispute will be decision that impacts, or *could* impact, the activity of motor systems. We can therefore restrict our empirical inquiry to cognitive/neural processes of decision that bear directly on whether motor pathways in the nervous system are activated. Notice that this can also include neural activity that *inhibits* action. Thus it does not matter whether the decision does in fact result in physical activity because decisions to abstain from action could also activate inhibitory pathways. The key here, to avoid stacking the deck in favor of one theory or the other is to emphasize *could*, not *do*.

This should follow from the observation that “judgment” absent some minimal connection to motor activation—i.e. judgment₁—leaves little to observe empirically, save for perhaps the introspective report of the research subjects.⁴³ When dealing with psychopaths introspective report becomes even more problematic than in the standard case because of the well-documented tendency of psychopathic individuals to disregard the literal truth (Hare 1993). But even if we were willing to take the subjects at their word, and even if every single one of them wanted to be truthful, we would in any case leave *them* no way to distinguish a genuine judgment₁ from a less certain or more ambiguous mental state or reaction in themselves. It is the connection to potential motor activity that really distinguishes the mental state whose causal powers Humeans and anti-Humeans want to understand.

Reaction time experiments in psychology provide a useful illustration of this point: we measure the time elapsed between presentation of a stimulus and a physical action such as a keystroke because we want to know roughly how long it takes a subject to form a judgment. Prior to the physical action, such as the subject’s executing a keystroke or initiating a verbal response, we have to assume that the judgment is incomplete. This is the innocuous sense in

---

⁴³ Nevertheless, some naturalists don’t seem too put off by the well-documented limitations of introspective report (e.g. Kennett 2002).
which I mean that judgment is a species of decision about what to do. As I hope to show, this distinguishing feature of genuine judgment can be made concrete and observable by appeal to anatomical considerations. Genuine judgments ought to implicate brain regions that contribute to networks that extend to motor pathways.

But it is not just any kind of motor activity-linked judgment that is at issue but rather *moral* judgment. This makes matters even more complicated since it requires that we have an operational definition of morality too. Though this is a difficult project in its own right, I think we can safely sketch some of the basic features of moral judgment.

By and large philosophers agree that moral concepts are normative. They are concepts with practical implications about what one ought to do. It is perhaps unnecessary to determine the origin of the norm in order to classify a claim or concept as normative. Some norms are social, some cultural, some personal, some religious, and so on. This is why moral psychologists have generally lumped moral concepts together with social concepts. For better or worse, then, if we want to look to the moral psychology literature we will probably have to accept this operational definition of morality. Psychologists and neuroscientists interested in moral cognition and behavior have consistently tied moral and social values together. It would be interesting to see whether we could observe differences that would allow us to dissociate social and moral norms or values for the purposes of empirical research, but I won’t pursue that project here.

As we turn to the empirical literature to see how the Humeanism dispute fares, then, moral judgments will have to be decisions about which actions or behaviors or responses to take—decisions about what it is best, or perhaps optimal, to do—in contexts of choice involving
norms or values. These operational definitions guide us to specific lines of empirical inquiry already underway.

Much work has already been done on the neural correlates of valuation, choice selection, and their relationship to motor pathways. The development of “neuroeconomics” (the intersection of neuroscience and economics) has given us an entire field devoted to understanding the neural mechanisms of decision-making in real subjects. But Humeanism and anti-Humeanism are not theories of decision-making in general but rather theories concerning the relationship between decision-making and motivation in moral contexts. Fortunately one of the most widely investigated branches of neuroscience is the neuroscience of motivation. In the remaining chapters I consider how Humeanism and anti-Humeanism fare in comparison to the data from neuroeconomics.

The issue before us is now whether the functional roles for belief and desire appear in the relevant capacity in the neuroscience of decision and motivation. But this brings us back to the beginning: how are we to recognize theoretical realizers for beliefs and desires qua folk roles when we see them? The entire dispute seems to hinge on how we answer the question. The functionalists who insist that it would be astounding if no such roles appeared in future neuroscience say virtually nothing about what these roles might look like there. Jackson and Pettit (1990) seem to have the idea that the expert on mental states, a cognitive scientist perhaps, will be the type of person who knows how to fill in the dots for an input condition like:

If there is a red box directly in front of x and..., then x will believe that there is a red box in front of x’ (17).

While the layperson might fill in the dots implicitly, the expert knows how to fill them in explicitly. This was Schiffer’s argument against the functionalist account of belief and desire
discussed in the previous chapter. While Jackson and Pettit disagree with Schiffer about whether
the layperson will be able to fill in the dots, all three seem to agree that the expert can.

This isn’t especially helpful because it tells us nothing about the nature of belief or desire. But then it isn’t intended to be. The nature of these states is characterized by folk *platitudes*. So it seems that as members of the folk we’ll have to just know them when we see them. Consequently, I’ll first review the empirical literature on the science of moral judgment and motivation and only then consider the most likely candidate for theoretical realizers of the folk roles in the story that emerges. Once we’ve found the most plausible candidate for theoretical realizers we will be in a position to determine whether the realizers serve in the capacity that we need them to in order to vindicate the disputes that depend upon them, such as the disputes about Humeanism and internalism.
5. Humeanism and the Neuroscience of Decision

§1. Preliminaries

This chapter begins an extensive review and assessment of current data on the neural mechanisms of decision. It will be no part of my argument to claim that the science of decision is complete, or that all or even many or most of the neural mechanisms involved in decision have been identified or studied, or even that the ones that have are beyond controversy. But a thorough review of the empirical literature should give us some idea of the direction that “future neuroscience” (as Jackson, Pettit and Smith among other philosophers sometimes call it) is headed. In particular, it should give us a better idea of the structure of neuroscientific explanations of moral phenomena like moral judgment and motivation. We need not have all of the details in place in order to get a sense for the basic explanatory framework. Once we look carefully at that framework we will be in a much better position to evaluate how Humeanism and anti-Humeanism fare as empirical theories of moral agency.

In this chapter I lay out some of the findings that have paved the way for the development of “neuroeconomics,” a relatively nascent research project situated at the confluence of economic theory and neuroscience. I then sketch some of the details of a new model of decision-making in the mammalian brain that has begun to emerge in that field. With the details of the decision model in place, I turn to the question of how the cellular and molecular neuroscience research that gave rise to the model intersects with higher-level research in the neural and cognitive sciences concerned with decision-making. My goal is to show that there is already in place a direct link between the data in cellular and molecular neuroscience, neurophysiology, neuroanatomy, pharmacology, clinical neuroscience, and ultimately cognitive neuroscience.
That link, as we shall see in this chapter, is perhaps strongest in the already massive body of literature on the neuroscience of emotion and the contribution of emotional processes to decision-making and motivation (§7). Then in the next and final chapter I will complete the empirical project, and hence the second horn of the dilemma, by showing how the model of decision-making presented here is directly relevant to decision-making in specifically moral contexts. The idea that we can already directly connect economic decision to emotion and the latter to moral and social decision, as we will see in the next chapter, shows that the sciences are now converging toward a model of decision that is immediately relevant to our philosophical theories of moral cognition, and moral decision and motivation in particular.

§2. Valuation and the Common Currency Hypothesis

Interest in the neurophysiology of decision-making has its origins in theoretical models of decision in economics. Traditionally, economists were not interested in the mechanisms or algorithms of decision-making but only in predicting and explaining what people would do with a range of options. As Milton Friedman (1953) puts it, traditional economists are interested only in “as if” models: models of behavior that view subjects as if they first construct a list of possible options and then select the highest ordered option among them. Unsurprisingly, traditionally minded economists (e.g., Milton 1953 and Gul and Pesendorfer 2008) argue that the neurophysiological correlates of these models are irrelevant; the concept of utility does not apply at the level of algorithm or implementation (Kable & Glimcher, henceforth K&G 2009). Still, many contemporary economists believe that uncovering the neural mechanisms of decision-making will lead to new and useful insights into primate behavior.
This is evidenced by the recent expansion of neuroeconomics, a field whose focus is the physiological mechanisms by which choices are generated and selected in actual subjects. Neuroeconomics is interested not in “as if” models, but in “because” models: neurophysiological models concerned with explaining behavior and its causes as well as the algorithm and implementation involved in decision-making. While “as if” models seek only to predict and explain a subject’s observable choices, “because” models seek to uncover the neurophysiologic constraints in place at the algorithmic and implementation levels, which can be used to validate behavioral predictions (K&G 2009).

Despite the contrast between these two types of models, the neurophysiological models share with the traditional models the idea that subjects represent a range of options on a single scale of desirability.44 Nearly all theories of decision over the years, including expected utility theory, prospect theory, and reinforcement learning theory, have been committed to the idea that subjects integrate the many dimensions of an option into a single subjective value. This value is the “common currency” for choice, which subjects determine before selecting the option with the greatest value in an aspect of decision-making known as “valuation.” The traditional models work only as if this were the case while the neurophysiological models have recently begun to vindicate the idea that this is actually the case, and so have retained the assumption (K&G 2009).

\[\text{\footnotesize 44 Notice the terminology: options are represented on a scale of desirability. This is an early appearance of familiar folk language in neuroscience research, and part of the reason that I think naturalists do better to concede full-blown eliminativism about folk mental states to the realist about beliefs and desires. Scientists sometimes talk this way too. Later in the chapter I will assess the extent to which we can rightly take this as evidence for theoretical realizers of the kind needed by the moral functionalists.}\]
§3. Valuation: Striatum, VMPFC, OFC, and the Midbrain Dopamine System

In particular, there is some evidence for subjective representation of value in the ventromedial prefrontal cortex (VMPFC) and striatum in the mammalian brain. The VMPFC is a part of the prefrontal cortex, the anterior part of the frontal lobe (i.e., the part that sits at the front of the brain toward the forehead). Scientists believe the prefrontal cortex to be involved in “executive function,” a blanket term for a variety of functions including goal-directed behavior, social control, planning, differentiating between stimuli, predicting future events and consequences, problem solving, and other kinds of complex thought. The VMPFC has been widely studied for its role in decision-making and emotion.

The striatum is a subcortical (located beneath the cerebral cortex) part of the cerebrum (i.e., the cortex, basal ganglia, and limbic system together). Specifically, it is one of the major parts of the basal ganglia system, a system of several strongly connected nuclei located at the base of the cerebrum. The striatum roughly consists of three parts, the caudate nucleus, putamen, and nucleus accumbens. It receives most of the input to the basal ganglia from other brain areas and sends inhibitory output to the output nuclei of the basal ganglia. Like the VMPFC, the striatum is involved in executive function and cognitive planning, though it is also involved in motor or movement planning.

Recently, both the VMPFC and the striatum have become the focus of a series of studies concerned with valuation in decision making. Researchers began to study valuation as an independent stage in the decision-making process in part because behavioral evidence with capuchin monkeys indicates that choice selection depends not on simple association resulting from trial and error but from more complex cognitive processes. Traditional theories of animal choice, such as those that invoke the concept of optimality (e.g. Kamil et al. 1987; Samuelson
1947) tend to employ Skinnerian (behaviorist) assumptions. They assume that prior experience gives rise to stimulus-response associations that the animal uses to select from its options. On this associative model, the values placed on the food options are not real psychological entities, though they appear to the external observer as though they are. Rather, they are simply associations between a particular stimulus (one of the foods) and a response (its selection).

Recently, however, Padoa-Schioppa et al. (2006a) tested this theory in a sequence of two experiments. In the first, they presented the capuchins with qualitatively different foods (e.g., a raisin and a slice of apple). In test sessions they determined which of the foods the monkey preferred, i.e., which it chose all else being equal. Then in a series of trials the monkeys were presented with choices between their preferred (A) and non-preferred (B) food, counterbalanced in left-side and right-side presentation, in varying quantities represented as ratios of A to B (e.g., 1B:2A, 1B:1A, 2B:1A, 3B:1A, 4B:1A, 6B:1A, etc.). Results indicated that monkeys always choose their preferred food when it is presented in ratios of 1:1 or better. However, when the non-preferred food was presented in significantly greater quantities than the preferred food the monkeys chose the non-preferred food instead. These results indicate that the monkeys take the foods to be comparable, and that they are comparable in their \textit{value}, a property consisting of both the physical attributes of the foods as well as the individual preferences of the monkey (Padoa-Schioppa et al. 2006a).

On the associative model of animal food choice, the monkeys’ preferences can be explained in terms of stimulus-response. Following the monkey’s initial choice, it begins to form an association between the choice offered (e.g., 4 pieces of food B to 1 piece of food A) and a particular selection (1A). Then in subsequent trials, the monkey is merely associating that particular pairing with a selection and responding to that association. A potential problem for
this model arises, however, when it comes to accounting for the monkey’s choice under novel circumstances. Can the associative model explain the monkey’s selection when from novel food pairings, ones that he has not yet had time to form associations about?

In a second experiment designed to test this question, the researchers presented the monkeys with food pairings that they had never previously encountered together. In familiarization sessions the monkeys chose between different amounts of novel foods. In subsequent test sessions they were offered choices between novel food pairings, where the pairings consisted either of one novel food and one familiar or two novel foods.

Results indicated that from the first trial monkeys were as effective in choosing between novel food pairings as they were familiar pairings. The associative model, in treating the monkey’s choice as a matter of stimulus-response association, predicts no relationship between the first and last trial in which a monkey is faced with a particular pairing. If the associative model is correct then the monkey’s choice in the first trial for that pairing should be random while the monkey’s last choice in the last trial for that pairing should be consistent, reflecting a learned association over the course of the trials. But in fact the researchers found just the opposite: the first trial choice and the last were identical 79% of the time, ruling out the possibility of fast stimulus-response association learning.

Padoa-Schioppa et al. take these results to tell against the associative model and in favor of a more complex cognitive model. If stimulus-response associations, which eschew the mental, fail to explain the monkey’s behavior, it is plausible that what is missing is the story about the relevant cognitive processes. The obvious alternative seems to be that the animal is cognitively representing the values of its options and then selecting the food with the highest
subjective value. Thus the researchers proposed a theory according to which economic choice involves a two-stage process of value-assignment and decision-making.

One consequence of the proposed two-stage cognitive model is that it makes the values real psychological entities. Whereas the traditional “as if” models make no claims about the algorithms of selection in real subjects, the “because” models are committed to the idea that these values are actually psychologically represented. Thus, if the model is correct, we should expect to find some physiological mechanisms of valuation, or value representation, at work during decision tasks. The question then becomes how primates actually represent these subjective values, if that is in fact what they are doing. These “because” models owe an account of the physiological mechanisms underlying the process of valuation. Where, and how, are these values represented in real brains? This brings us to the VMPFC and striatum.

Emerging evidence suggests that representation of value takes place, at least in large part, in these two areas of the mammalian brain. Some pioneering work with the role of the striatum in decision-making by Samejima and colleagues (2005) examined primate estimations of reward for choices in decision-making tasks. This research was based on similar reinforcement learning models of decision-making according to which organisms adaptively choose actions in three steps: (1) they estimate the value of an action, (2) they then compare the values and select an action based on the outcome, and (3) they update the action values using the errors of their estimated action values (Samejima et al. 2005; Sutton & Barto 1998). This hypothesized sequence has received support in the literature.

Reinforcement learning models of the basal ganglia (of which, recall, the striatum is the major input nuclei) have been established in both animal models (e.g., Ito & Doya 2009) and in

---

45 Though there are many variations on reinforcement learning (RL) models, generally all center on these three steps (Daw & Doya 2006).
human subjects (e.g., O’Doherty et al. 2004). One of the pillars of these experimental paradigms is the well-established function of midbrain dopamine neurons. The midbrain, a major part of the brainstem, is made up of the tectum on its dorsal surface (the roof of the midbrain), and the cerebral peduncle, which includes the tegmentum and substantia nigra.

The midbrain contains bundles of cells, or nuclei, that connect the components of the motor systems. In particular, the ventral tegmental area (VTA), i.e. the ventral part of the tegmentum, and the substantia nigra, which is part of the basal ganglia, are essentially the points of origin for the dopamine cells that make up the brain’s dopamine system (the mesocorticolimbic dopamine system). The midbrain dopamine system is well known for its role in the production of movement; midbrain dopamine neurons are implicated in movement disorders such as Parkinson’s disease. The striatum is the major target of the signals from these dopaminergic neurons.

These dopaminergic neurons have been shown to encode changes, or both positive and negative errors, in expectations of reward in primates (Morris et al. 2006; Schultz et al. 1997). Dopaminergic neurons of the VTA and the substantia nigra transmit signals to the striatum, nucleus accumbens, and frontal cortex, which are structures known to be involved in motivation and goal-directed behavior. The signals transmitted by these neurons are short, phasic bursts activated by the presence of desirable or appetitive stimuli, such as food or juice in the case of primate experiments. When the activity of these dopaminergic neurons is recorded in monkeys engaged in experimental tasks for which they are rewarded, phasic dopamine activation can be seen to occur in response to the so-called appetitive event, e.g., the monkey’s touching a piece of apple or tasting juice (Schultz et al. 1997, Figure 1).
In one type of task, monkeys are trained to pull a lever for a reward in response to the appearance of a visual cue, such as a flash of light. Initially, the phasic dopamine activation occurs upon presentation of the reward. When the monkeys are repeatedly presented with the cue prior to reward presentation, the dopamine activation eventually shifts and begins to occur in response to the onset of the cue rather than the presentation of the reward, just as in classical conditioning paradigms. Furthermore, when the reward is not delivered following the onset of the cue to monkeys trained to expect the reward following the onset of the cue, the rate of activation in the dopamine neurons becomes depressed below the base rate of activation at precisely the time the reward should have occurred. This indicates that the dopamine activity actually encodes the expected time of reward delivery or error in the animal’s predictions about the occurrence of appetitive events (Schultz et al. 1997).

Crucially, these dopamine neurons encode error in reward expectation rather than merely report the occurrence of reward because they respond to the difference between the predicted time and magnitude of the reward and the actual time and magnitude of the reward. This is sometimes called the reward-prediction error. Spiking, or phasic neuronal firing, increases when the reward (or appetitive event) is better than predicted, it remains unchanged when the reward occurs just as predicted, and it decreases when the reward is worse than predicted (Schultz et al. 1997). Similar results have been obtained in experiments concerned with motivational function in decision-making and learning.

Dopamine neurons in monkeys similarly code for reward expectation errors (that is, errors in the monkeys’ expectations of reward) during decision tasks, or tasks in which particular decisions lead to particular rewards with scheduled probabilities (Satoh et al. 2003). In an experiment of this type, Samejima and colleagues (2005) recorded the activity of striatal neurons
in monkeys engaged in reward-based decision-making tasks. Monkeys were trained to turn a handle to the left or to the right for rewards. Over five blocks of trials the probability of receiving a large reward rather than a small reward for a left or right turn was varied. For example, in the first block of trials the monkey would receive a large reward for a left turn with 90% probability and a large reward for right turn with 50% probability (i.e. a 90-50 trial block). The five blocks used were 90-50, 50-90, 50-10, 10-50, and 50-50.

The crucial feature of this experimental design is that it allowed the researchers to distinguish between neuronal activity associated with the monkey’s recognition of the value of an action (“action value”) and neuronal activity associated with the action chosen (“action choice”). So, for example, even though the monkey should prefer the left turn in both 90-50 and 50-10 trial blocks, the researchers could determine whether the neurons encoded a change in action value (i.e., the fact that the monkey is aware of the change in value even though in both cases it prefers the left turn and so takes that action).

The researchers recorded the activity of 504 striatal projection neurons in the right putamen and caudate nucleus. Two crucial points emerged from their results. First, some of these neurons in fact appear to encode action value rather than action choice. In comparing the discharge rates of neurons from asymmetrically rewarded blocks of trials, the researchers found that a single neuron with a significantly higher discharge rate in a 90-50 trial block than in a 10-50 block (i.e. in asymmetric blocks) did not have a significantly different discharge rate between 50-10 and 50-90 blocks, for which the preferred actions are opposite. This shows that the neuron encodes the left action value (i.e., the 50% probability), not the action itself. If the neuron were recording the action, it would show a significantly different discharge rate across these blocks for which the preferred actions are opposite. A similar pattern was found for right action values, and
even for the difference between left and right turn action values, suggesting that some of the neurons code specifically for these values. In total, roughly one third of all the striatal neurons (43/142) were found to record action value.

Second, the results also indicate that the activity of the neurons encoding action value predicted the monkey’s choice of action. Using the past action of a monkey in previous trials (i.e. which handle turn the monkey chose) and the past reward (i.e. whether it received a large or small reward for that turn) in a reinforcement-learning model, the researchers were able to successfully predict the probability of the monkey’s subsequent action choices (Samejima et al. 2005, Fig. 4A).

These results indicate that a portion of dopaminergic neurons in the striatum represents action values, that those values are updated based upon previous experience of action and reward, and that they determine future action. That researchers can successfully predict future action under a reinforcement learning algorithm using estimated reward values suggests that these striatal neurons encoding action values play a crucial role in the process of selecting an action when faced with a decision.

Several lines of research have implicated the ventromedial prefrontal cortex (VMPFC) in decision-making and its intersection with emotion. Just as the striatum appears to be the locus for representations of action values, which figure into action selection, the VMPFC and

46 The scientific literature is often unclear about the difference between the orbitofrontal cortex (OFC) and the ventromedial prefrontal cortex (VMPFC). This is in part because functional differences have not yet become well established. Both the OFC and VMPFC are regions of the frontal cortex implicated in emotion. References to the VMPFC are either broad or narrow. In the broad sense, the VMPFC is a region containing the OFC in its ventral-most (lowest) part. In the narrow sense, the VMPFC is the area immediately superior to (just above) the medial OFC. Where possible I will specify which of these regions is under consideration. In most of the instances to follow, however, the term is intended in the broader sense. In these cases the term is intended to include rather than exclude the OFC.
orbitofrontal cortex (OFC) appear to be the locus for representation of the subjective values of different rewards. Previously I discussed the work of Padoa-Schioppa and Assad (2006, 2008) in generating evidence for cognitive models of decision-making as opposed to traditional stimulus-response conditioning. Their behavioral research with monkeys showed that the animals were subjectively representing the values of food choices. But they also investigated the neurological basis for this subjective value representation.

Padoa-Schioppa and Assad (2006) showed that neurons in the OFC encode the values of both offered goods and chosen goods. They presented monkeys with a series of trials in which they could choose between pairs of juices. In each trial they varied the amount and type of juice offered. Recording the choices, the researchers were able to calculate the subjective value of each juice option for the monkey on a common scale. This allowed them to make hypotheses concerning the monkey’s neuronal representation of the values of each of the juices. They could then check the neuronal activity against these behavioral hypotheses.

The results identified three distinct neuronal patterns corresponding to three types of neuronal function. A portion of the 931 cells in the OFC from which recordings were taken showed a firing rate significantly correlated with the subjective values previously hypothesized from the behavioral data. These neurons therefore seem to be offer value neurons because they track the subjective value of the juice option offered. In other words, the activity of the particular neuron co-varied with the value of the juice on offer.

A second subset of the neurons showed a firing rate linearly correlated ($R^2 = .86$) with the subjective value of the juice (i.e. reward) that the monkey actually chose (or will eventually choose). In this case the neuronal activity was low when the monkey chose the juice with a chosen value score of about 2, higher when it chose a juice with a value score of about 4, and
highest when the monkey chose a juice with a value score of about 6.\textsuperscript{47} That these variations in cell activity are significantly correlated indicates that they represent the subjective value of the chosen reward. This subset of neurons was therefore labeled chosen value neurons.

The third subset of neurons showed a distinct categorical or binary firing activity response to particular juices. The researchers labeled these taste neurons. Accordingly, each of these three classes showed a distinct timing pattern. Offer value and chosen value neurons predominantly fired immediately following presentation of juice options while taste neurons fired after the juice reward was presented (Padoa-Schioppa & Assad 2006; K&G 2009).

These same researchers (Podoa-Schioppa & Assad 2008) later extended the investigation and showed that these neuronal value representations were “menu-invariant.” That is, the neural responses are representations of the economic value of individual goods rather than representations of relative value, or the value of a good relative to its paired alternative. Recording from 557 individual neurons in the OFC, the researchers presented the monkeys with competing juice pairs (i.e. offers). To determine whether the neuronal responses depended upon menu (i.e., what alternatives are available at that particular time), they recorded the neuronal activity while the monkeys chose between three different juices (A, B, and C in decreasing order of preference) in varying amounts, presented in interleaved pairings of A:B, B:C, and C:A.

The results again showed three patterns of neuronal activity corresponding to three types of neurons (offer value, chosen value, and taste neurons) as discussed above. Moreover, the results showed that the neuronal responses were in fact invariant for changes of menu. The

\textsuperscript{47} Value scores for the rewards (i.e. juices) are calculated as discussed above. For example, call water “juice A” and unsweetened Kool-Aid “juice B.” One portion of B versus two portions of A is represented: 1B:2A. Offer types range from 1B:2A to 10B:1A. If behavioral evidence shows that the monkey chooses A when paired with 1B, 2B, and 3B, it is indifferent at 4B:1A, and it chooses B when 6B and 10B are offered, then the value of 1A is roughly equal to the value of 4B [i.e.: V(1A)=V(4.1B)] and hence has a value of approximately 4.
neuronal activity encoding the value of one of the juices was largely independent of availability of other juices. This is especially interesting in light of behavioral data for transitivity. In behavioral experiments (as described above) the monkeys’ choices exhibited transitivity. That is, monkeys who preferred juice A to juice B and juice B to juice C preferred A to C (Padoa-Schioppa & Assad 2008; K&G 2009). Establishing menu invariance of neuronal activity takes on additional importance then because it shows that the neuronal responses, like the behavioral responses, were stable and consistent, and therefore reflect transitivity (Padoa-Schioppa & Assad 2008 Figure 3A).

Transitivity is crucial because the concept of economic value depends upon it. If a subject prefers A to B and B to C, then she must prefer A to C if the economic value of each item (A, B, and C) is to be preserved. Thus, evidence of transitivity in neuronal activity supports the idea that the values of goods are represented in a common, comparable currency in the OFC neurons. In other words, transitivity is only possible if the neurons encode individual subjective values of goods on a single, common scale and not merely relative (menu-variant) values. Each good on offer, then, has its absolute subjective value represented by particular neurons on a common scale.

Importantly, these results have support in the literature. Lau and Glimcher (2008) produced results similar to those of Padoa-Schioppa and Assad (2006, 2008) for neuronal activity in the striatum. In an oculomotor choice task with monkeys, they found distinct task-related responses in phasically activated striatal neurons (PANs) corresponding to the three types of neurons distinguished in the study by Padoa-Schioppa and Assad: action value neurons, chosen value neurons, and choice neurons, where the last correspond to taste neurons. The time frames for neuronal responses were very similar (Lau and Glimcher 2008; K&G 2009). This
evidence for transitivity is crucial to the received model of decision in neuroeconomics, which I review shortly. Up to this point I have focused on the first stage in that model of decision-making: valuation. I turn now to the development of the second stage: choice.


These findings concerning the encoding of absolute subjective values in the VMPFC, OFC, and striatum, which therefore make transitivity possible in decision-tasks, can be linked to recent research on the role of neurons in the lateral prefrontal and parietal cortex in choice selection. Up to this point the research under consideration has been concerned with the first stage in models of decision-making: valuation. The second stage in such models is choice, or the subject’s selection from among the evaluated options and subsequent implementation.

Investigation of the neural correlates of choice selection and implementation in decision-making tasks has implicated neurons in the lateral prefrontal cortex and parietal cortex in these processes. Much of this research centers on the visuo-saccadic control system of the monkey, which is made up of the lateral intraparietal area (LIP), the frontal eyefields (FEF), and the superior colliculus (SC) (Glimcher 2003; K&G 2009, Figure 7). This system is known to control visual saccades, or rapid movements of the eye between different points of fixation. It has also been studied for its role in perceptual decision-making.

The connection between visual saccades and decision-making is quite complex. The basic idea underlying this research is that considering one’s options involves representation of the external world based on perceptual stimuli, and these stimuli affect pathways that control action, such as reaching, shifts in attention, and especially eye movement. In other words,
decision-making involves a connection between gathering visual data and executing future saccades toward the appropriate environmental objects (Glimcher 2001).

In many ways the connection between an animal’s control of visual saccades and its ability to make choices is, despite its complexity, rather unsurprising. Since at least as far back as Phineas Gage, researchers have observed that individuals with damage to the prefrontal cortex exhibit deficits in planning and organizing behavior, memory, and problem solving. Decision tasks require some or all of these functions for successful performance. Subjects must consider their options, keep information concerning each of those options “online” for further processing, plan and anticipate the consequences of their choices, and constantly update their information stores as they acquire experience with the task. From there it is a relatively short step to understanding how the saccadic control system is involved in decision-making. One of the crucial components of decision-making is holding information online for processing, which is the key feature of the working memory system.

Researchers have known that the frontal lobe is an important part of the working memory system since as far back as the 1930s when the delayed-response task was first developed. In this task, monkeys were shown food contained in one of two identical holes in a table. Identical covers were then placed over the holes, time was elapsed, and the monkeys were presented with the choice between the holes. Monkeys with damage to their prefrontal lobes consistently performed much worse in the task than monkeys without such damage. Moreover, performance in the task decreases as the length of the delay period increases (Bear et al. 2001: 769). This suggests that the frontal lobe is a crucial part of the working memory system.

More recently these findings have been substantiated by research with human subjects engaged in the Wisconsin card-sorting task, which requires them to sort cards according to an
unstated sorting rule. Subjects begin by sorting however they like, but through successive updates about sorting errors they quickly learn the implied rule according to which their sorting is being judged. After ten correct placements, the rule is changed and subjects must adjust accordingly. Subjects with normal brains have relatively little difficulty adjusting to the rule change, while those with damage to the prefrontal cortex struggle to adjust to the rule changes. Typically, they continue to sort cards according to the previous rule. Success in the task requires holding information about previous cards and errors in working memory in order to use that memory to alter future choices. Thus it seems that patients with damage to the prefrontal cortex are unable to use recently acquired information to update their choices, which once again implicates prefrontal cortex in working memory (Bear et al. 2001).

Researchers have since shown that other areas of cortex are involved in working memory or the storage of information, particularly some areas of the parietal lobe, which is directly posterior to the frontal lobe. The parietal lobe includes the posterior parietal cortex (BA 5), known to integrate sensory and motor information, the primary somatosensory cortex (areas 3, 1, and 2) known to be involved in processing sensory information, and area 7, which is posterior to the somatosensory cortex and known to be involved in vision and proprioception (i.e., feedback about the status or location of our body’s parts in space), specifically locating objects spatially.48

These parts of the parietal lobe are essential parts of the motor planning system. They are extensively connected with the frontal lobe, which as we have seen is known to be involved in decision-making and planning, and they are also extensively connected to area 6, which is made up of the premotor cortex (or the premotor area or PMA) and the supplementary motor area (SMA).

---

48 BA 7 in the parietal cortex will re-emerge later on in this chapter, as it has been implicated in non-moral practical judgments in fMRI research.
This is important for two reasons. First, area 6 is now well known to play a central role in planning complex movement. For example, Roland and colleagues used emission tomography (PET)—a form of brain scan that monitors increases in cortical activation by detecting the gamma rays emitted by a biologically active tracer artificially introduced into the body—to measure changes in cortical activation as subjects mentally rehearsed finger movements without physically moving their fingers. They found that area 6 was active under such conditions but not area 4, the primary motor cortex known to be involved in the execution of movement (Roland et al. 1980, 1996). In other words, the areas of the parietal lobe that are a part of the working memory system are extensively connected to the part of the brain implicated in planning complex movement.

The second reason this is important is that area 6 is one of two major sources (the other is area 4, the primary motor cortex) of axons that descend to the corticospinal tract, the pathway that connects the higher-order cognitive areas of the brain that deal with decisions and planning and the intention to move to the spinal cord where those intentions are carried out. These points are important because they suggest a direct structural or anatomical connection between brain areas implicated in working memory, higher-level motor activity (such as motor planning and coordination) and decision-making, which in turn supports the idea that these functions are directly linked as well.

One of these areas of the parietal lobe that has recently received much attention for its role in working memory and decision-making is the lateral intraparietal cortex (area LIP), located in the intraparietal sulcus on the surface of the parietal lobe. Area LIP is best known for its role in the guidance of visual saccades, or quick eye movements that occur regularly. Briefly, the mammalian visual system essentially works by building up visual images from a collection of
data, which the eyes gather by oscillating about quickly, executing saccades, and bringing various parts of the visual field into greater focus. Electrical stimulation of area LIP is known to produce these visual saccades.

Moreover, in the late 1980s, research with macaque monkeys in a delayed-saccade task showed that some neurons in area LIP exhibit a particular response property connected to working memory. In a delayed-saccade task, monkeys are trained to fixate on a central fixation target on a computer screen. They must maintain their gaze on the fixation point during the brief appearance and subsequent disappearance of a target stimulus in the periphery and move their gaze to the point in the periphery where the target had appeared following a set delay period. Neurons in area LIP (the activity of 141 cells was recorded, 24 of which were considered “intended movement cells” on the basis of preliminary data which recorded their responses to visual activity) begin firing right after the presentation of the peripheral target. But crucially they continue firing during the delay period after the target has disappeared and until the saccadic eye movement takes place. Further research showed that many of these cells were active even when a visual stimulus never fell within their visual response field. This suggests that the activity of the cells corresponds to the pending saccade rather than the retinotopic location of the visual stimulus on the retina (Gnadt & Anderson 1988; Bear et al. 2001).

So began a research project connecting the control of visual saccades to the working memory system in the mammalian brain. In addition to area LIP, the saccadic control system (also known as the visuo-saccadic control system or visuo-saccadic decision system) includes the frontal eye fields (FEF) and the superior colliculus. The FEF are located in the premotor cortex in area 6, which sits roughly between areas 4 and 8. The superior colliculus (SC) is located
atop the midbrain. The SC is composed of alternating layers of cells and axons (Kandel et al. 2000). The superficial (i.e. uppermost or surface) layers receive projections from retinal ganglion cells creating a topographic map of the visual field. Neural activity at a particular point on the topographic map results in shifts of attention or action toward the point in space corresponding to that point. Both area LIP and FEF also have topographic maps, similar in function to the topographic map of the SC. Area LIP projects to both (Glimcher 2009: 516). Moreover, these areas are reciprocally connected, which makes it possible for them to function as a network. Ultimately the SC is connected to the brainstem circuits that control the eye movements.

The line of research under consideration here essentially began with the discovery that the neurons in the SC of a monkey that is presented with two saccadic targets equal in subjective value become active in the two locations on the topographic map representing the amplitude and direction of those saccades. Glimcher and Sparks (1992) found that when one of the two saccadic targets was identified as more valuable by the monkey a burst of activity occurred at the collicular site corresponding to the more valuable target and activity the alternative collicular site was suppressed. Subsequent research revealed that neural firing rates in the SC increase and decrease along with reward probability in a graded manner, which indicates that the desirability or subjective value of each of the possible saccades as based on reward probability and magnitude is encoded by neural firing rates in the SC (Dorris & Munoz 1998; Glimcher and Sparks 1992; Platt and Glimcher 1999; K&G 2009).

---

49 The superior colliculus (SC) is the homolog in mammals of the optic tectum in other vertebrates. Because the experiments under consideration here were performed with mammals, I refer to the SC, though it is common for researchers to refer to the tectum, especially to “tectal maps,” which are just the topographic collicular maps discussed here.
Given these findings, Platt and Glimcher (1999) decided to test whether these neural activations preceding saccades were encoding subjective values of *movements*. They focused on area LIP, which is upstream (i.e. prior in the sequence of signal transduction) from the colliculus. As we have seen, area LIP is involved in the production of visual saccades and also directly linked to working memory. They recorded from LIP neurons while manipulating the probability of receiving a reward for a given saccade or the magnitude of the reward. They found that the rate of firing in area LIP was almost a linear function of reward probability and magnitude (K&G 2009; Platt & Glimcher 1999).

This research suggests that choice selection from among one’s options amounts to a neural process in which the saccade associated with the highest rate of neural firing is selected and implemented in motor pathways downstream. This accords well with the anatomical considerations concerning area LIP with which we began this section.

Importantly, however, the value representations of the saccades in these fronto-parietal maps differ from the representations of subjective value in the striatum, VMPFC, and OFC neurons discussed above in that they are not menu invariant. Rather, these values are *relative* to the values of the other saccade possibilities (Dorris and Glimcher 2004; K&G 2009, Figure 8). That is, when researchers tracked the firing rates of the area LIP neurons encoding reward values associated with saccades, they found that the firing rates increased for larger rewards as expected, but did not increase when all of the reward magnitudes were doubled. If the neuronal representations of value in area LIP were tracking absolute rather than relative subjective values like the neurons in the striatum, the firing rate should increase along with increased overall reward magnitude. Instead, researchers found no increase, which suggests that these parietal
neurons encode rescaled values. Kable and Glimcher (2009) suggest that these neurons rescale the absolute values in order to maximize the differences between options.

A concurrent line of research has implicated these fronto-parietal networks in related perceptual decision-making tasks. In perceptual decision-making tasks, primates are trained to fixate their gaze on a central fixation point on a computer screen. Two targets then appear, one to the left periphery and one to the right. This is done so that one of the two targets appears in the response field (RF) of the neurons from which activity is recorded and one outside the RF. Next, a group of moving dots appears between the two targets in what’s known as a random-dot motion pattern. The direction and motion of the dots is randomly chosen on each trial. Following the offset of the dots, a delay period is imposed before monkeys execute a visual saccade to either the left or the right to indicate the direction of motion in the dot pattern. The researchers recorded the activity of single neurons in area LIP during the task that exhibited spatially selective persistent activity. In other words, the neurons began firing in response to the appearance of the saccade target and maintained a steady level of discharge through a delay period and until the saccade was executed.

Moreover, these neurons are actively only before an eye movement is made, which indicates that they are responsive to space during the delay period and not to the subsequent eye movement (Shadlen et al. 1996; Shadlen & Newsome 2001). Neurons in area LIP that exhibit this pattern are of particular interest because they are active during the crucial decision period, which occurs between the initial sensory response to the visual target and the motor activity that produces the saccade.

The researchers found that these neurons changed their rate of firing in a way that predicted the monkey’s saccadic eye movement. The neurons became active during the period in
which the monkey viewed the moving dot pattern, and became more reliable (i.e. the predictability of the chosen saccade based on prior neural activity improves) as the viewing time increased (Shadlen & Newsome 2001, Figure 10, p. 1925). Additionally, the strength (coherent movement of the dots) and motion (amount of movement) of the pattern affected the timing and magnitude of the neural response. Stronger directions of motion (more dots in coherent motion and greater movement) yielded larger neural responses earlier in the viewing period when the stimulus was presented in the RF. Alternatively, when the signal was presented outside the RF (i.e., in this case to the left) the neural activity was suppressed (S&N Figure 9 p. 1925).

Roitman and Shadlen (2002) later repeated the experiment with one important variation: they trained the monkeys to respond as soon as they had made a decision about the motion of the dots rather than to wait for the set time interval to expire. Reaction time measurements were taken in order to determine the amount of time it took for the monkey to reach a decision. This variation allowed the researchers to determine the actual moment at which the monkey committed to a behavioral response. Thus they were able to determine that the activity of the neurons in area LIP that accompanied the motion-viewing period was not a consequence of the decision itself. Also, the addition of reaction time data made it possible to test whether the amount of time it took the monkey to reach a decision was related to the rate of growth and decline in the neural activity. They found an inverse relationship between the spiking rate of the LIP neurons and the reaction time across trials for all recorded neurons (p < .0001).

Perhaps the most important result of the study, however, was the finding that the activity in LIP during the decision-making period leading up to the ~50 millisecond period between the decision and execution of choice was ramped. The activity gradually increased until a value threshold was reached. This suggests that neural activity prior to the decision involves the
accumulation of information about the motion of the dots until it reaches a threshold value at which point the monkey initiates the saccade approximately 50 milliseconds later (Roitman and Shadlen 2002, Figure 7).

These results mesh with an account of the neural mechanisms responsible for producing visual saccades that has since achieved near consensus among researchers. As we saw previously, area LIP, the FEF, and SC form a network with reciprocal connections. The SC ultimately connects with the brainstem circuits that initiate eye movements. This connection is mediated by a particular class of collicular neurons, known as “burst” neurons, that fire action potentials in both a low frequency state characterized by many different firing rates and in a burst state characterized by a fixed, high rate of firing. It is now widely agreed that eye movement generation requires that the burst neurons achieve a firing-rate above a particular threshold, at which point a self-perpetuating burst begins and persists until the eye movement has been executed (Glimcher 2009: 516).

Coupled with the data concerning valuation discussed previously, this research concerning eye-movement generation within the context of choice yields a comprehensive two-stage model of decision-making in the mammalian brain. I therefore conclude this section with a summary account of that model—based upon the research so far presented and some additional considerations from economic theory and neuroeconomics—and then turn in the next section to convergence with recent human subjects research concerning decision, especially in social and moral contexts.
§5. A Summary Model of Decision-Making in Neuroeconomics

The empirical considerations so far considered are converging toward a unified model of decision-making. Paul Glimcher has been at the forefront of that project. He has recently provided a few comprehensive reviews of the data from neuroeconomics that have made such an account possible (see Kable & Glimcher 2009 & Glimcher 2009). Here I review the central elements of that emerging model of decision-making, based on the data presented thus far, that I think have immediate implications for the Humeanism dispute. It bears repeating that neuroeconomics is a relatively nascent discipline and that some of the mechanisms discussed here are likely to be revised as the data proliferates. Still, several of the most important elements of the model are already well established and we can expect to get from this review a reasonably clear idea of the direction that the “future neuroscience” of decision is heading.

The neural mechanisms of decision-making in primates unfold in two stages: valuation and choice. Valuation is the process whereby all available goods or actions in a decision context are ascribed a subjective value—encoded in the firing rates of neurons—which are based on the individual’s subjective preferences in conjunction with the objective features of the good or action, all of which is learned in part through previous experience; choice is the process whereby information concerning the option with the highest subjective value is implemented in the neural circuitry that ultimately produces physical movement.

The fundamental piece of evidence for the first stage, valuation, is the observation that levels of neural activation in the striatum and frontal cortex are linearly correlated with subjective values as reported by the subjects, particularly in the form of selections in behavioral tasks. I previously reviewed several lines of investigation, especially research with primates, which converge on this conclusion. In other words, the populations of neurons in particular
regions of the brain actually encode the subjective values of goods or actions prior to choice (and in fact even in contexts in which no choice is actually made, e.g., Delgado et al. 2000). Neurons in these areas of the brain, then, both learn and represent the values of goods or actions.

Moreover, we have seen that these values are represented on a common, single scale. For complex reasons having to do with the limitations of expected utility theory in neuroeconomics, that scale is not a scale of utility, or a scale on which the points correspond to units of well-being. Values in the context of choice are not utilities. Instead, the values—called subjective values (SVs)—are real numbers where the units are, given the data under consideration, action potentials per second.

One crucial consequence of taking the units of value to be action potentials per second rather than utilities is that the model can accommodate the fact that an individual’s choices do not always relate to her well being or utility. The present model of decision can accommodate the fact that people sometimes make bad choices, or at least choices that would not be dictated by the rational calculus of utility theory – choices that do not leave them better off. And this, it seems clear, they regularly do (Kahneman & Tversky 1979).

Moreover, because SVs are always consistent with choice while utility is not, using SVs as the units of valuation allows us to predict violations of the dictates of expected utility theory. Somewhat crudely, if the relevant populations of cells in your brain dramatically increase their rate of firing for a particular option in the process of choice, even when selecting that particular option will not actually increase your overall well being, then there remains an explanation for why you chose it anyway. Neurons with rates of firing linearly proportional to utility would by definition obey the axioms of expected utility theory, and this would make it impossible for us to understand real choices (Glimcher 2009: 508).
Subjective values are thus the mean firing rates of particular populations of neurons in response to goods or actions, and in this sense they are cardinal (i.e. they denote quantity as opposed to just an ordinal position or rank). For an object that can become options in the context of choice, its subjective value is capable of being encoded in the mechanisms constitutive of valuation in the brain. This yields the result that “subjective values” are neural representations of both absolute and relative features of external objects. This raises an important philosophical question: what is it out there in the world that these SVs represent? We have seen already that the short answer to this question is both the subjective preferences and objective features of the items or actions on offer. But what exactly are these?

We have only briefly touched on the longer answer to this question. Subjective values are learned through experience. This learning process involves the phasic activation of dopamine neurons rooted in the midbrain, in particular, neurons originating from the substantia nigra pars compacta (SNC) and the ventral tegmental area (VTA) and projecting to the basal ganglia and frontal cortex. In some of the initial work on valuation processes with monkeys, Schultz et al. (1993) showed that individual dopamine neurons in these midbrain dopamine pathways responded strongly to unconditioned rewards (i.e. rewards that produce a response even in monkeys not previously trained to respond to the stimulus) but not to conditioned rewards. This was some of the first evidence that these dopamine neurons encode something other than merely pleasurable experiences. The later work by Montague et al. summarized previously established the idea that what those neurons actually encode is the difference between the monkeys’ expectations about the reward and its actual experience of an obtained reward. This is the reward-prediction error. It is important because it helps to explain the portion of subjective values based on individual, subjective preferences.
To say that the subjective values of items or actions are partly determined by individual preferences is to say that those values are *partly* determined by the individual’s experience of reward, or perhaps more simply, that individual’s response to that reward. Neuroeconomists call that experience *ExperiencedR*. *ExperiencedR* is the experience of a positive outcome from a reward. Rewards are objects or events that generate a particular class of behaviors (such as consummatory and approach behavior) that support basic and essential processes. Thus the experience of reward not only produces such behaviors but also positively engages with emotion systems and influences learning processes (Glimcher 2009).

The other half of subjective values, the part that represents objective features of objects or actions in the world, is expected reward or *ExpectedR*. *ExpectedR* is essentially a prediction. It is a prediction about the magnitude and probability of a reward given the previously accumulated evidence. That evidence takes the form of stored information about the physical, motivational, and economic properties of an item or action. The information is in an important sense objective: it concerns the actual physical properties of items or actions and the actual physical effects of those properties on the organism. This is one sense in which subjective values are objective.

It turns out, of course, that even this is a bit oversimplified. Both the objectivity and subjectivity of subjective values are intertwined in *both* expected and experienced reward. To put it simply, we might think of the *objectivity* of subjective values arising from the impact of physical and chemical characteristics of objects that impact upon our sensory organs and the *subjectivity* of subjective values as arising from the experience produced by the processing of that electrochemical information in our nervous systems. Because both the interplay between our sensory organs and the physical world *and* our experience of the changes effected in the
nervous system figure into both expected and experienced reward, it turns out that objectivity and subjectivity figure into subjective values twice, once in the construction of ExperiencedR and once in the construction of ExpectedR. Neuroeconomists define dopamine firing rates as:

\[ DA(\text{spikes/s}) = \alpha(\text{ExperiencedR} - \text{ExpectedR}) \]

where \( \alpha \) is a scaling parameter that controls the rate of learning.\(^{50}\) The crucial point for the Humeanism dispute is that objectivity and subjectivity figure into both parts of the formal definition of subjective values in the neural model (I will return to this important point in the final chapter).

Those subjective values are learned through experience. The learning takes place in networks distributed in the basal ganglia and frontal cortex. More specifically, we have seen that the neuronal networks implicated in the learning of values based on reward-prediction errors have their origins in the striatum, the ventral striatum in particular, and project to the frontal cortex, and the medial prefrontal cortex in particular. Much is already known about the low-level mechanisms that constitute learning in these areas. The learning involved in valuation is made possible by synaptic plasticity, or changes in the strength of the connections between neurons. The networks of dopamine neurons disseminate information concerning learned subjective values throughout these regions of the brain.

\(^{50}\) The idea behind the scaling parameter for learning rate, \( \alpha \), is basically as follows. Economic models tell us that reward-prediction errors (the difference between expectation and actual reward) can't change, or scale, with the variance in the set of choices on offer, because then the subjective value of a good or action would be dependent upon the choice set, and that in turn would make transitivity impossible. Recall that transitivity is a crucial feature of the models under consideration – it has been repeatedly demonstrated that human and animal choices obey transitivity. However, the empirical data indicate that the reward-prediction error signal in dopamine neurons is variance dependent. Neuroeconomists have resolved this problem by determining that it is the learning rate—the rate at which the differences between reward prediction and experience are learned—that scales with variance in choice sets rather than reward-prediction error.
All of this, together with the data previously reviewed, suggests that subjective values are encoded in the medial prefrontal cortex and ventral striatum of the mammalian brain. The data suggests that these areas serve as the point at which subjective values are implemented into the mechanisms that produce choice, the second stage in the two-stage model.

Research with the saccadic control system, the FEF, area LIP, and SC in particular, indicates that the subjective value representations (SVs), which become relative subjective values (RSVs) in the context of a choice set, are distributed by the network connecting the ventral striatum and medial prefrontal cortex. These RSVs drive the firing rates of collicular burst neurons via activation of the cells in the FEF-LIP-SC network. Recall that the cells in these areas form topographic maps of objects in the visual world and of the eye movements required to direct gaze to those objects. The value representations, encoded in neural firing rates, propagate through the network. When one of the value representations—in the form of a neural signal—drives the collicular burst neurons associated with a particular location on the topographic map above its firing-rate threshold, a self-perpetuating burst occurs that results in activation of cells in the anatomically connected brain stem systems that in turn produce movement of the eyes to the associated location in space. This, ultimately, is choice selection.

As we move to the research conducted with human subjects it will be helpful to lay the groundwork here for the connection I am proposing. At present one of the crucial questions that this decision model raises for future research is whether and how the model applies to choice sets containing more abstract objects. This is particularly important for my purposes because presumably the objects or actions we choose in moral decisions have abstract properties, or properties that at least appear to depend more on ideas and norms and values than physical or
chemical properties. The property of moral rightness, for example, seems to be more abstract than the physical properties of a food or juice reward.

This seems to be the main line of objection that most philosophers have to the idea that neuroeconomics will reveal anything of interest about moral cognition and motivation. Though it is a reasonable concern, I think there are several convincing lines of reply available. Perhaps the most important among them is that objecting in this way seems to overlook the point of the discussion in Chapter 1, which is to make clear that the core disputes about moral cognition in metaethics presuppose the existence of beliefs and desires. That ultimately is as I see it the single most important empirical assumption that gets imported by the traditional metaethical framework, and that is ultimately the assumption that I will challenge directly.

It is especially important to keep in mind that the Humeanism dispute, which is the real focus of this dissertation, is presented as a dispute about the *standard picture of psychology*, not just a standard dispute about *moral* psychology. It is a picture that Smith tells us we inherit from Hume, whose writings were not exclusively concerned with moral psychology. As I discuss in Chapter 1, what gets the fundamental dilemma of metaethics off the ground at all is the distinction between beliefs and desires, which is made explicit in proposition (3), the Humean theory of motivation, with the clause: “where beliefs and desires are distinct existences.”

That aside, many will insist that the data discussed here, collected using monkeys, grapes, and raisins, has little or no relevance to more abstract choices and judgments, like moral judgments. While I have no doubt that a neural account that can adequately accommodate such abstract goods is a long way off, what the objection lacks is any convincing reason to suppose that non-moral and less-abstract psychological processes will differ from moral and abstract psychological processes in *kind* rather than merely in degree. However more complex moral
decisions may be than, say, economic decisions, it will be far more surprising if the neural explanations for moral decisions turn out to look nothing like the neural explanations for economic decision than that the two will resemble each other enough to justify membership in the same general category (namely the category of decisions). For example, the kinds of explanations for choice selection under consideration point directly to cellular responses to environmental changes, such as the propagation of signals at the cellular level from motor planning areas to descending motor pathways. There is no obvious reason to expect that neural explanations for more abstract decisions (including perhaps moral decisions) will require appeal to an entirely different kind of mechanism. It’s hardly clear what that alternative mechanism could even be. In the simplest terms, any influence that abstractness, even in the form of moral normativity, exerts on decision-making at the neural level will likely have to utilize some or most of the mechanisms under discussion here, and so there is no obvious reason to doubt continuity between economic decision and more abstract decision. It certainly isn’t sufficient to insist that they just seem different in kind. That is why I set out to challenge the “standard picture” of human psychology adopted by so many moral philosophers rather than some other picture of moral psychology.

And that is also why I think it is important, as these empirical chapters reveal, to investigate the similarities and differences between psychological and neural processes in social and moral contexts and those processes in economic contexts. That these processes both require a contribution from pathways and structures associated with emotion, for example, or that the same structures and pathways are consistently implicated in both contexts, suggests that we were on the right track in beginning with the hypothesis that cognition and moral cognition differ in
degree rather than kind. It was an empirical hypothesis, and perhaps a controversial one, but one that may now be finding some empirical support.

It is also revealing, I think, that philosophers inclined to object to the relevance of the present research on the basis of a deep distinction between the abstract and the concrete or the moral and the non-moral aren’t more explicit in claiming that moral psychology differs in kind and not merely degree from non-moral psychology in their own traditional research projects. The objection only works, if it works, in both directions. But few philosophical papers on the Humeanism dispute begin with disclaimers about the deep distinction between moral psychology and regular psychology.

Similarly, there is no good reason to think that such an objection presents a special problem for the application of neuroeconomics and neuroscience to moral philosophy. If the objection really picked up on such a crucial distinction between the moral and non-moral or the abstract and concrete then it would certainly warrant as much concern for proponents of the traditional moral problem: “Surely belief-desire psychology is of a different kind of psychology than that involved in making moral judgments. Surely that framework is too simplistic. Judgments about what it is morally right to do are considerably more abstract and complex than simple beliefs and desires about tables and chairs, foods and rewards, and so on.” The fact that objections of this sort aren’t widely expressed suggests to me that few philosophers would really want to deny that moral cognition and motivation are entirely unlike other (standard, non-moral, concrete rather than abstract) kinds of cognition and motivation.

Finally, in the interest of giving an empirical reply to the objection, it is also worth noting that a few recent studies suggest that the neural networks involved in saccadic choice tasks have access to neural information that maps abstract properties to actions. For example, when
monkeys are made to choose between two different color targets in the visual field that constantly switch locations, the neurons in the SC show activity that reflects instantaneous mapping of color and value, even when the association is changed from trial to trial (Sugrue et al. 2004; Horwitz et al. 2004; Gold & Shadlen 2000). In other words, neurons in the SC must have access to information that relates the abstract properties of the choices to actions (Glimcher 2009). Though abstract properties of saccade choices like color and location in one’s visual field are admittedly a long way from moral properties, these studies are an important first step toward understanding how the choice selection mechanisms under discussion apply to choice sets where the items or actions have more abstract properties, such as a changing color or location.

Research up to this point has yielded two models of the choice selection mechanism, one for reaction-time tasks and one for non-reaction time economic choices. For the former, Shadlen and colleagues (Roitman & Shadlen 2002; Gold and Shadlen 2000; Shadlen and Newsome 2001) have proposed a “race-to-barrier” model according to which a choice is immediately made when the value of one option exceeds a set threshold of neural activation. For the latter, researchers have proposed the previously discussed “winner-takes-all” model in which the option with the highest value is selected (e.g. Dorris & Glimcher 2004; K&G 2009). Crucially, Wang and colleagues (Lo and Wang 2006; Wang 2008; Wong and Wang 2006) have now demonstrated that both methods of choice—threshold-based selection and decision-rule selection—can be implemented in a single circuit (see especially Wang 2008 for discussion).

These models are similar, and what is important here is that they describe the same mechanism: the FEF-area LIP-SC network receives information concerning the subjective values of the available options, and then relates those representations on a single scale in order to maximize the differences between them before selection. The subjective values are stochastic
because they are quantities of spikes, which in turn dependent upon the strengths of synaptic connections. The selection, mirroring higher-level data (e.g., Kahneman & Tversky 1979), is stochastic. In other words, at the neural level the mechanisms of choice are as we would expect from higher-level empirical data concerning behavior: statistically predictable but not precisely. Choice in the mammalian brain, then, is an imperfect process. Neurons encode subjective values, not utility, and consequently choice sometimes violates the axioms of expected utility theory.

At present, one of the best explanations for this phenomenon has to do with the effects that other neural networks and structures have on the networks that connect the frontal cortex and striatum. One of the best candidate neural systems for this task is, of course, the emotion system. As Glimcher says, these violations of expected utility theory “doubtless reflect the influence of emotion-related brain structures on medial prefrontal cortical and ventral striatal activity” (2009: 519). I turn now to research with human subjects that support the present model of decision-making. We need to determine inasmuch as it is possible whether the model of decision under consideration is applicable to human subjects, and more specifically, human moral agents.

§6. Convergence with Human Subjects Research

Evidence for the role of the striatum and VMPFC in encoding subjective value is not limited to primate research. Functional magnetic resonance imaging (fMRI) of the brains of humans engaged in various reward-based decision-making tasks has produced data convergent with the above primate research at a higher level of neural function.
fMRI offers one of the few relatively noninvasive looks into the brains of living human beings, and for this reason continues to grow in popularity despite its increasingly well-documented limitations (e.g., see Logothetis 2008 for a thorough review of fMRI, including a detailed account of what it is, how it works, and what we can and cannot expect to do with it in cognitive neuroscience, and Nichols & Newsome 1999 for a shorter and less technical account of the technology and some of its limitations).

Very roughly, fMRI is one method of measuring contrast in levels of blood oxygenation in the brain. In other words, fMRI of the sort under discussion here utilizes a blood-oxygen-level-dependent (BOLD) signal, a measure of differences in the amount of oxygen in blood. Hemoglobin is a protein in the blood that transports oxygen for use by cells. The physical properties of hemoglobin are such that oxygenated blood has little impact on an applied magnetic field (i.e. it is diamagnetic) while deoxygenated blood enhances an applied magnetic field (i.e. it is paramagnetic). Neural activation causes an increase in cerebral blood flow and glucose consumption and ultimately lowers the concentration of deoxyhemoglobin (or hemoglobin desaturated with oxygen molecules), which in turn increases the strength of BOLD signal. Researchers use this BOLD contrast measure in conjunction with statistical techniques to determine which regions of the brain are most active (i.e. reliably exhibit an increase in the strength of BOLD signal) during a particular task. The spatial resolution of the brain maps generated by fMRI is generally about 1 millimeter or less, while the temporal resolution is much less impressive, on the order of seconds.

I mention some of these basic principles here because they remind us to interpret fMRI research with caution. The technology as well as the physical processes and properties of nervous systems that make fMRI possible are astoundingly complex and not yet fully
understood. For example, while the BOLD contrast signal is correlated with changes in cerebral blood flow, the relationship between regional blood flow and the activity of individual neurons is still under investigation and at present remains a matter of controversy (Raichle and Mintun 2006). Moreover, the reliability of fMRI research hinges on several assumptions, some involving biological processes, some the technology and others involving the relationship between mind and brain. As an example of the latter, our ability to correlate the activity of various brain regions with mental tasks or functions depends whether the mind is really “modular,” that is whether it is really divisible into discrete parts according to function. If not, it will be impossible to correlate discrete mental functions or modules with the activity of particular brain regions.

Despite its limitations, the addition of fMRI to the neuroscientist’s toolkit has expanded our understanding of neural function and its relationship to cognition and behavior in many important ways. We should exercise some skepticism and caution when we look at the colorful images that result from fMRI research; but just the same, we should not ignore the potential uses of carefully crafted studies that employ the technology. fMRI data can be particularly helpful when coupled with other experimental techniques such as EEG, single cell recording and behavioral measures. These contrasting approaches are already beginning to converge on some of the central mechanisms of decision.

First, Plassmann et al. (2007) scanned the brains of hungry human subjects as they bid on various foods where only foods actually won could be eaten in the thirty minutes following the experiment. The auctioning procedure was designed to incentivize subjects to bid only the amount they were willing to pay (WTP) for the food item. This design allowed researchers to compare the results of the scans to a behavioral measure of subjective value, i.e., a measure of
their WTP as based on how much they actually bid. The results revealed a correlation between WTP, the measure of subjective value of the food item, and BOLD activity in the medial orbitofrontal cortex (mOFC) (Plassmann et al. 2007, Figure 3A, B, C). These results have since been replicated (e.g. Hare et al. 2008 and 2009; K&G 2009).

Second, Tom et al. (2007) isolated brain activity associated with the evaluation of gambles (i.e. decision utility) where subjects knew that the gambles would not be immediately resolved. This allowed them to distinguish the neural activity associated with the evaluation of potential outcomes from activity associated with experienced outcomes. Subjects were made to choose between accepting and rejecting various gambles what carried a 50/50 chance of gaining or losing money. Additionally, they were asked to reveal the strength or certainty of their decisions by labeling each acceptance or rejection as strong or weak. These reports were used to make a behavioral measure of sensitivity to gains and losses.

The results of brain scans revealed that activity of the VMPFC and dorsal and ventral striatum both increased with the amount of gain in the gamble and decreased with the amount of the loss. Interestingly, the results also indicated that a subject’s having greater behavioral loss aversion (i.e., behavioral indications of aversion to losing money in gambles) was associated with greater neural sensitivity to both losses and gains. Further investigation of the regions involved showed that for most participants the VMPFC and striatum activity had the following pattern: the slope of the decrease in activity for increasing losses was greater than the slope of the increase in activity for increasing gains (Tom et al. 2007). In other words, results revealed a distinct pattern of neural loss aversion.

Perhaps one of the most interesting features of the research by Tom and colleagues is that the experimental design allows for a distinction between the emotional component of loss and
the decision-making component. Whereas previous research has tended to focus on the increased activation of brain regions associated with negative emotional reactions to experienced or anticipated losses (e.g. amygdala and anterior insula), this study suggests that potential losses during decision-making tasks are represented by a decrease in neural activity associated with the encoding of the subjective value of goods (Tom et al. 2007: 517).

This is important given the previously discussed primate research because that research indicates that individual dopamine neurons in the VMPFC and striatum actually encode subjective value. This is a remarkable point of convergence between single cell (“lower-level”) primate research and BOLD activity measure (“higher-level”) research with human subjects implicating the VMPFC and striatum in valuation during decision-making.

§7. Decision Making and Human Emotion

This last study raises questions about the role of emotion in decision-making. Research with human subjects over the past decade has provided substantial gains in understanding the role of emotion in judgment and decision, including judgment and decision in specifically moral contexts. In this section I explain in general terms the development of emotion research that led scientists to focus extensively on the ventromedial prefrontal cortex (VMPFC), since today the VMPFC remains at the center of research concerned with the interaction of decision-making and emotion, which is in turn crucial to understanding contemporary work on social and moral judgment and decision.

The key points that I want to establish in the discussion to follow are (1) that emotion makes a crucial contribution to decision-making and thus bears directly on the decision research from neuroeconomics discussed above; (2) that on the basis of the extensive convergence of data
ranging from the cognitive to the molecular—data that implicates the same neural networks and pathways—a great deal is now known about the specific neural mechanisms by which that contribution is made.

In defending these claims I lay the groundwork for the thesis of the final chapter and the overarching theme of the positive project of this dissertation. I will argue that this extensive convergence of data provides a compelling argument for the idea that the model of decision emerging from recent work in neuroeconomics as discussed above likely has immediate relevance to philosophical disputes about moral cognition and motivation given the contribution of emotion to moral and social-decision making. As we will see in this section, moral and social decisions depend upon a crucial contribution from emotion systems. Because the same structures, networks and pathways are implicated in economic decision-making are implicated in specifically social and moral decision-making, it is emotion that provides the needed link between the neuroscience of decision, especially neuroeconomics, and contemporary philosophical disputes about moral judgment, decision and motivation.

Emotion is a broad concept with slightly different connotations in different research circles, so it will useful here to provide some background on recent research into the role of emotion in decision-making processes. One of the leading figures in this area is Antonio Damasio. Since the early 1990s, Damasio has championed the “somatic marker hypothesis” (henceforth “SMH”) according to which emotions play a crucial role in decision-making processes (e.g., Damasio et al. 1991; Damsio 1994, 1996, & 1999). His research has generally centered on VMPFC and the amygdala, since both brain structures have been previously implicated in emotion. We can begin with his account of emotion:

An emotion is defined as a collection of changes in body and brain states triggered by a dedicated brain system that responds to specific contents of one’s perceptions, actual or
recalled, relative to a particular object or event… The responses toward the body proper enacted in a body (somatic) state involve physiological modifications. These modifications range from changes in internal milieu and viscera that may not be perceptible to an external observer (e.g., endocrine release, heart rate, smooth muscle contraction) to changes in the musculoskeletal system that may be obvious to an external observer (e.g., posture, facial expression, specific behaviors such as freezing, flight and fight, and so on) (Bechara & Damasio 2005: 339-40).

Specifically, these responses bring about the release of certain neurotransmitters, active modification of somatosensory maps, and modification of transmission signals from the body to somatosensory regions. This “ensemble” of enacted body-brain responses constitutes an emotion. The feeling is the phenomenon perceptible to the organism in which these responses are enacted. It is made up of the ensemble of signals mapped in somatosensory regions of the brain. And while Damasio uses the term emotion to refer to the ensemble of enacted responses, he uses the term somatic state to pick out the body-related responses characteristic of an emotion.

Damasio’s account is one of several competing theories of emotion (e.g., the James-Lange theory, Cannon-Bard theory, Schachter theory, Arnold theory, etc.) The main point of contention among these theories concerns the relationship between cognitive and physiological states (Kandel et al. 2000). So while there is some dispute about which states are tied up with the definition of emotion (e.g., the Arnold theory denies that autonomic responses are an essential part of the definition), it is probably fair to say that most of the competing theories recognize some role for the cognitive and physiological states contained in Damasio’s definition.51

51 Damasio’s account is an updated version of the James-Lange theory, as is Schachter’s.
Damasio divides emotions into primary and secondary emotions. The primary emotions are the somatic reactions to various stimuli and features of stimuli present in the world or in our own bodies, which we have from very early in life. Some examples might be the feeling of fear you experience in response to the movement of an animal, like the slithering of a snake, or the fear you experience in response to your own bodily configuration or state, such as seeing your finger sliced open and bleeding. These are primary in the sense that they are innate and pre-organized. One need not even recognize the object or stimulus causing the feeling to experience the emotion. Children need not understand what exactly a snake is in order to recoil at its slithering.

Secondary emotions, or what he sometimes calls “adult emotions,” are much less primal and much more intellectual. Unlike preorganized reactions to snakes, secondary emotion begins with the processing of mental imagery and lead to nonconscious, involuntary somatic responses to those images. For example, in hearing of the death of a close friend, you might form mental images of that person, your relationship with him or her, possible scenarios that led to his or her death, and so forth. You then experience involuntary somatic states in response to those images.

Many of the “emotions” under consideration in this chapter concern this second variety. Participants in gambling tasks, for example, experience somatic, emotional states in response to mental images and predictions about the prospects of financial gain and loss. In much the same way, research subjects faced with moral choices like whether to flip a switch to change the course of a runaway trolley in order to save five people at the cost of one engage with mental

---

52 Damasio makes the distinction between primary and secondary emotions in *Descartes’ Error* (1994). Elsewhere he distinguishes between primary and secondary inducers of somatic states (see Bechara & Damasio 2005). See footnote 52 for further discussion.
images and cognitive reflection about possible outcomes. It is these mental images that activate the somatic states.

Damasio’s distinction between primary and secondary emotions may correspond roughly to the distinction between what other scientists sometimes call emotional states (or just “emotion”) and conscious emotional feelings (or just “feeling”). Emotional states involve activation of peripheral and autonomic nervous systems, the endocrine system and the skeletomotor system as well as subcortical structures like the amygdala, hypothalamus and brain stem. Feeling by contrast involves activation of the cerebral cortex, cingulate cortex and the frontal lobes. These neural networks are distinct but extensively connected (Kandel et al. 2000). Fear, for example, involves both physiological responses, such as dry mouth and increased heart rate, and cognitive responses like our awareness of these physiological responses and conscious

53 Damsio’s terminology, particularly his definitions of “emotion” and “feeling,” are not orthodox (see 1994: 146), which can complicate discussion of his work in conjunction with work conducted by other scientists. Damasio distinguishes not only between emotions (roughly somatic states) and feelings (roughly the perception of somatic states), but also primary emotions (preorganized, Jamesian, visceral emotions developed very early in life) and secondary emotions (the kind of emotion triggered cognitively, such as by imagining or remembering, which requires developing connections between categories of objects and situations and primary emotions; see 1994: 131-37) and primary inducers (innate or learned stimuli that elicit a somatic state) and secondary inducers (entities generated by recall of a primary inducer) of somatic states (Bechara & Damasio 2005). So far as I can tell, there is little difference between a primary emotion and a somatic state induced by a primary inducer, so it isn’t exactly clear why Damasio shifted the discussion from primary/secondary emotions (1994) to inducers (2005). However, that shift seems to be present also in The Feeling of What Happens (1999). There he emphasizes that we must keep in mind that both inducers (environmental stimuli) and emotions have ranges, which helps to account for variation in emotion across people and cultures (see pp. 56-9). Thus, maintaining the distinction may help to promote precision and specificity in emotion research. In any case the crucial point here is that, terminological differences aside, both Damasio and other scientists recognize a distinction of sorts between the more primitive or visceral somatic states associated with limbic system/amygdala/anterior cingulate activation on one hand and the more associative, cognitive, imagination- and memory-based somatic states associated with additional activation of somatosensory and prefrontal cortices on the other hand.
understanding that we are afraid. The extent to which these networks are interconnected is evidenced in part by the evolving and expanding definition of the “limbic system.”

The concept of a limbic “system” began with Broca’s (1878) observation that a group of cortical areas bordering the brain stem and corpus callosum (the bundle of fibers connecting the brain’s two hemispheres) on the medial walls of the mammalian brain. He named these distinctive cortical regions the “limbic lobe” because *limbus* is Latin for “border.” Then in the 1930s the American neurologist James Papez suggested that the medial wall of the brain contains an emotion “system” which connects the cortex to the hypothalamus. The group of structures he included in the proposed emotion system—including the hippocampus, hypothalamus, anterior thalamic nuclei, cingulate cortex, fornix and neocortex—came to be called the Papez circuit. By the 1950s, Paul MacLean, an American physiologist, had popularized the term “limbic system,” which he used to refer to the emotional processing network designated by both Broca’s limbic lobe and Papez’s circuit. In short, the network of structures involved in emotion that border the brain stem came to be known as the limbic system.

Neuroscientists now recognize that there are several difficulties with treating this interconnected network as a discrete system. The concept of a “system” suggests that the structures are united by a single, common function. But as we have already seen there is plenty of evidence to suggest that there is more than one neural system responsible for the production and experience of emotion. Moreover, many of the structures included in the system are known to serve additional functions beyond the production or processing of emotion. For at least these reasons it is now widely recognized that the limbic system is not really a single system with a singular unified function. Nevertheless, researchers continue to use the term to refer to the network of structures involved in emotional processing. And consequently the limbic system
has grown to include additional structures and areas, particularly the amygdala and the OFC (Bear et al. 2001; Kandel et al. 2000).

In these terms, Damasio’s crucial contribution to our understanding of emotion has been his emphasis on the possibility that conscious emotional feeling is the brain’s interpretation of somatic or bodily responses. As we shall see, his research broadly supports a dissociation between these two networks because, like Phineas Gage, the subjects in his studies with frontal lobe damage have normal somatic responses to startle stimuli yet systematically fail to generate those somatic responses for cognitive representations of future consequences.

Moral philosophers, it seems to me, use the term “emotion” broadly to refer to both of these networks, and sometimes fail to specify which is under consideration. This suggests once again that the use of data from neuroscience to solve traditional philosophical problems is a considerably more complicated matter than many philosophers would have us believe. Moral philosophers who agree that emotion is crucial to moral psychology, for example, may turn out to have entirely different ideas, that is, different from each other and different from emotion researchers, about what exactly that means. In any case, the data to follow is directly concerned with feeling and indirectly with the “visceral” states from which feeling arises. The brain regions under consideration are evidence of this – they are by and large cortical rather than subcortical structures, though of course the anatomical connections between these will be important.

Damasio’s inspiration for the SMH largely came out of clinical work with patient EVR, who underwent a surgical treatment for meningioma at the age of 35. The surgery resulted in bilateral ablation of his VMPFC. Much like the famous Phineas Gage, EVR’s frontal lobe damage substantially altered his behavior and personality. Once a successful, intelligent, and
hard-working individual, EVR became incapable of planning, unable to maintain a steady job, and unable to make decisions efficiently about matters ranging from the mundane to the important. Following the surgery, EVR fit the clinical diagnosis for acquired (as opposed to developmental) sociopathy, though he retained his superior intelligence, verbal abilities, and knowledge of social norms and so forth (Damasio et al. 1991).

A series of behavioral studies and interviews revealed that EVR’s defective planning and decision-making was the result not of a loss of social knowledge but rather the ability to select the most advantageous response from among his response options in the course of a decision. In Damasio’s technical language, EVR and EVR-like patients are still capable of accessing knowledge about both “manifest” meanings of situations, i.e. basic knowledge about the identity of a person or a place or situation, and “implied” meanings of situations, i.e. the positive or negative value of a person, place, or situation, the possible response options in a situation, and even the predicted consequences of each of those options. Despite access to this knowledge patients with damage like EVR’s systematically fail to choose the most valuable or advantageous option in response to relatively complex social situations, where implied meanings are crucial. These observations led Damasio and colleagues to propose a mechanism for deficits of this kind, namely, a deficit in the activation of somatic markers (126).

According to Damasio, when normal individuals make decisions they do so in a way that falls somewhere in between conscious deliberation and automatic or instinctive selection. For most of us, this process is aided by somatic markers, which are essentially bodily or affective signals that mark the predicted consequences of our selections with a negative or positive valence. I consider the consequences of a potential action by imagining the future event. Perhaps I find that it is accompanied by a bad visceral, “gut” reaction, which draws my attention
to the negative consequences of the option. This is the hypothesized behavioral-level operation of a somatic marker. But there is a less obvious and equally important hypothesized function of these somatic markers: modulation of neural systems such as the serotonin and dopamine specific systems that engage with the activation and suppression of appetitive and aversive behaviors. In other words, even in the absence of conscious experience the somatic markers may activate inhibitory or excitatory neural systems in order to decrease or increase the likelihood of a particular response.

According to the SMH, then, somatic markers act in at least these two capacities to make it possible for us to select from among a daunting set of response options in complex situations, particularly social situations where the importance of implied meanings increase the complexity of decision-making. This hypothesis gains some traction in cognitive models of information integration for the purposes of representation and recollection. One of the classical problems for researchers who want to understand the relationship between the mind, and in particular conscious experience, and the brain is the so-called “binding problem,” the problem of determining how and where the brain integrates a very large and diverse set of perceptual information for the purpose of representation and recollection.

To make the point simply, it seems clear that diverse sensory modalities such as vision, smell, taste, and so forth must come together to accurately represent an experience of the world or some particular object or situation. And though different brain regions and mechanisms have been implicated in each of these distinct sensory modalities, the question remains how or where all of the information is integrated so as to produce an accurate representation. Scientists working on this problem have generally thought that this sensory integration takes place in so-
called multimodal cortices, portions of the anterior temporal and anterior frontal regions of the brain, where a final multimodal representation is formed.

For a variety of complex reasons Damasio and colleagues eventually rejected this solution in favor of an alternative model according to which the “binding” of a representation involves phase-locked co-activation of distinct sites of neural activity in sensory and motor cortices (1989: 127-8). In other words, there is no single unified brain area for the integration of fragmented sensory and motor information into a single representation. Rather the information is integrated from many distinct regions connected by so-called convergence zones, or regions which synchronize patterns of neural activity corresponding to fragments of the representation that were topographically organized during the initial experience of the object or event. And this temporal integration, according to Damasio and colleagues, takes place in the ventromedial prefrontal cortex.

The idea was that convergence zones located in the VMPFC keep a record (so to speak) of simultaneously active regions. Moreover, this temporal record of activation would need to retain the information through delay periods in order for decisions to be possible. The VMPFC was considered a good candidate for this role because it is well established that the frontal cortex is required for successful execution of any task in which a delay period precedes the response. That is, the frontal cortex is the locus in the mammalian brain of the ability to hold information online for processing (Fuster 1973, 1989). Of course, Damasio’s particular binding hypothesis—and in fact the binding problem itself—is controversial and not a hypothesis I intend to defend. I mention it here only because it helps to explain some of what motivated Damasio’s focus on the VMPFC and the relevance of his distinction between implied and manifest meanings at the
neural level. It also helps to contextualize the development of Damsio’s SMH as an explanation for the deficits exhibited by patients like EVR with frontal lobe damage.

In complex social decisions, such as those faced by EVR, implied meanings abound. In deciding how to handle oneself in a heated dispute, for example, considerations might include the number of witnesses, the loyalties of those witnesses, the predicted reaction of the individuals involved, the proximity of escape routes, previous experiences with like scenarios, and so on. Each of these components contributes to what Damasio calls the implied meaning of the option, and many of these components will involve different representations from distinct brain regions, which must be activated simultaneously and held “online” for processing. Thus complex social decisions will require substantial synchronization of cognitive components, each of which must be attended to if a decision is to be made effectively (1989, 1991).

Somatic markers, then, increase the likelihood that the outcome of this complex task is an advantageous selection by labeling the response options with somatic signals, essentially bodily feelings that occur consciously and subconsciously in the manner described above, to increase the efficiency and accuracy of the process. Because of the time-locked retroactivation theory of binding as well as the established functions of the frontal cortex as a locus for emotion, executive processes and working memory, this labeling was thought to take place in the VMPFC.

To test these hypotheses Damasio and colleagues (1991) set out to determine whether patients with frontal lobe damage like EVR’s (1) could in fact access knowledge of implied meanings and (2) would fail to activate somatic states in response to those implied meanings. Electrodermal skin conductance responses (SCRs) were used to measure activity of the autonomic nervous system in response to visual stimuli. SCR (also known as “galvanic skin response”) is a widely used measure of electrical resistance of the skin, which changes in
response to emotional arousal. Three groups of subjects were tested: five “normals” with no brain damage; five “bifrontals,” i.e. individuals with lesions to the orbital and lower mesial frontal regions of the brain; and six brain-damaged controls, i.e. subjects with brain lesions to regions outside the ventromedial frontal cortices (1991: 127).

SCR was measured in response to three types of stimuli: elementary unconditioned stimuli, such as a loud clap, which reliably elicit SCRs in normal subjects; target pictures with strong “implied” meanings, such as social disasters, mutilation, or nudity, which regularly elicit high-amplitude SCRs in normals; and non-target pictures, such as bland images of scenery and abstract patterns, which do not have implied meanings and do not elicit high-amplitude SCRs in normals. Presentations of 40 pictures (10 target and 30 non-target) were ordered randomly and presented to subjects in two conditions. In the passive condition, subjects simply viewed the slides. In the active condition, they were asked to comment on the content of the picture and its impact upon them.

Three points of interest emerged from comparison of SCRs across groups (Damasio et al. 1991: 129): (i) there were no differences across groups in response to elementary unconditioned stimuli; (ii) normals and brain-damaged controls showed significantly larger responses to targets than non-targets in both passive and active conditions; (iii) in the passive condition the responses of bifrontals to target pictures did not differ significantly from responses to non-targets (p = .5, Sign Test), while in the active condition their responses to targets were significantly larger (p = 0.031).

These results appear to support the SMH because they indicate abnormal passive autonomic responses to target stimuli from subjects with bifrontal damage. In other words, the bifrontals failed to generate the same SCRs to target stimuli that the subjects in the other two
groups did. Moreover, this finding cannot be explained by basic dysfunction of the autonomic nervous system since bifrontals showed normal SCRs to elementary unconditioned stimuli. From this, Damasio and colleagues concluded that bifrontals failed to activate the usual somatic state in response to target stimuli. Interestingly, when the subjects were required to comment, normal responses to target stimuli occurred. Thus it seems bifrontal subjects are still able to generate autonomic responses, and this suggests that it is the triggering mechanism that has been compromised (130).

Moreover, Damasio et al. argue that the anatomical arrangement of the ventromedial cortices, the areas damaged in the bifrontal patients, supports the hypothesis concerning the existence of the convergence zones discussed previously. These cortices receive projections from all sensory modalities either directly or indirectly, they are the only frontal regions known to project to central autonomic control areas, and they have an abundance of bidirectional connections with the hippocampus and amygdala (1991: 131). These latter structures are implicated in memory and emotion respectively.

The hippocampus, a seahorse-shaped structure situated inside the medial temporal lobe, is the part of the limbic system widely thought to be involved in the consolidation of long-term memory from short-term memory. The amygdala is an almond shape structure that sits just below the cortex on the medial side of the temporal lobe (bilaterally). It is usually considered part of the limbic system. Afferent connections to the amygdala come from all of the sensory systems, and interconnections within the amygdala integrate that sensory information. The amygdala has efferent connections with several structures including the hypothalamus, the reticular nucleus of the thalamus, and the ventral tegmental area (VTA). In particular, Damasio has elsewhere suggested that the amygdala is the structure responsible for inducing somatic
responses to directly experienced ("primary") environmental stimuli (2005: 340). I will return to discuss further research on the amygdala shortly. Here the crucial point is that Damasio took these anatomical considerations—the extensive connections between the amygdala, hippocampus and ventromedial cortices—to add additional support to the SMH given the hypothesized role of the VMPFC in synchronizing neural information for executive control and thus in facilitating somatic markers.

Following these initial experimental results, Damasio and colleagues began testing the SMH in experimental paradigms focused on decision processes specifically. Hence the link between VMPFC research, emotion research, and decision-making. In particular, they began a series of experiments with the Iowa Gambling Task (IGT), a task designed to simulate decision-making in real life circumstances by including elements of uncertainty, reward, and punishment. In this task, participants select cards from four decks (A through D), identical in appearance, over the course of 100 trials in an effort to acquire as much money as possible. Participants begin with $2,000. Each card choice results in financial reward or a combination of reward and penalty, which seems variable to the participant, though in fact the actual rewards and punishments have been fixed by the experimenters in advance. They receive $100 for each selection from decks A and B, and $50 for each selection from C and D. On deck A, five in ten selections incur a penalty between $35 and $150; on deck B, one in ten incurs a penalty of $1,250; on C, five in ten selections incur a penalty between $25 and $75; and on D, one in ten incurs a penalty of $250. Thus the high reward decks (A and B) incur greater amounts of punishment, specifically a net penalty of $250 every ten trials, than the lower reward decks (C and D), which yield a net profit of $250 with every ten selections. Therefore, because decks A
and B are disadvantageous while C and D are advantageous, successful performance (i.e. overall gain) requires selecting more from decks C and D than A and B.

The key feature of the IGT is that the participants have no way of knowing when penalties will arise. This, researchers claim, leads them to rely more on “hunches” or intuition than purely calculated decision (Bechara et al. 1996). Participants are told that the objective is to obtain as much money as possible in an unknown number of trials until they are told to stop (though in fact it is limited to exactly 100 trials), that the penalty schedule is unpredictable (though in fact it is fixed), that they will find that some decks are worse than others and that they may switch decks whenever they wish.

Patients with bilateral damage to VMPFC, i.e. EVR-like patients, consistently perform worse in the IGT than do healthy, normal participants. Normal controls quickly learn to select from the advantageous decks and to avoid the disadvantageous decks while the damaged patients do not. The differences between normals and EVR-type participants in card selection between decks are statistically significant. While both groups begin by sampling from all four decks, normals begin to favor C and D, returning to the high-risk/high-reward decks A and B only occasionally. EVR-like patients return more frequently to the disadvantageous decks (Bechara et al. 1994). Damasio and colleagues hypothesized that this difference in performance could be due to somatic marker deficits.

As in the target pictures paradigm, Damasio and colleagues used SCR measurements to detect autonomic activation—or somatic state activation—in response to the consequences of card selections. Experiments were performed with seven patients with bilateral damage to the VMPFC and 12 normal controls. Consistent with the previous study, they found that both groups generated SCRs in response to the selection of cards yielding penalties or rewards. After
some experience with the trials, however, normal controls began generating SCRs in anticipation of some selections.

Specifically, normals began generating anticipatory SCRs for the high-risk (disadvantageous) decks A and B, which were higher than anticipatory responses to C and D, while the subjects with VMPFC damage entirely failed to generate anticipatory responses. These results could not be explained by a general inability to produce SCRs since both groups produced SCRs in response to rewards and punishments. Nor could motion artifacts, i.e. that SCRs were generated by the movement of the hand toward a card, explain the results since (a) the SCRs were generated before any movement was initiated and (b) no anticipatory SCRs were generated in early trials before participants gained experience with the task and (c) recordings were taken from the non-moving hand. Moreover, SCR magnitude was not equal for high (A and B) and low (C and D) risk decks, and the magnitude of anticipatory SCRs increased from early selections from A and B to later selections, which indicates that the SCRs were related to increased experiences with the consequences of selection.

An alternative explanation, however, was that these subjects were simply less sensitive to reward and punishment than normals. To rule this out, Damasio and colleagues repeated the experiments with two modified versions of the task. In the first, they reversed the reward/punishment schedule such that successful performance was determined by differences in reward rather than differences punishment. Selection from advantageous decks resulted in greater immediate punishment but still greater delayed reward while selections from disadvantageous decks resulted in lower immediate punishment but lower future reward. Even in this altered task, VMPFC-damaged patients showed significantly poorer performance. This suggests that the cause of poor performance is not an impaired sensitivity to punishment or
reward, since an impairment of that sort would presumably result in more selections from the advantageous decks (in which immediate punishment was greater) and subsequently better overall performance.

In the second modified version, the future consequences of selection from the risky, disadvantageous decks were made more averse by increasing delayed punishment and decreasing delayed reward in those decks. If the cause of poor performance in brain-damaged subjects were due to insensitivity to consequences, then this modification should result in normalized performance. Results revealed that it did not: despite the modification, VMPFC-damaged subjects again performed significantly worse than healthy controls (Bechara et al. 2000; Dunn et al. 2006).

Given these findings, Damasio and colleagues concluded that patients with bilateral lesions to VMPFC failed to generate somatic markers for imagined future options based on previous experience with consequences, which hindered their ability to avoid the disadvantageous decks and therefore their overall performance in the task. Moreover, the results given the modifications to the task indicate that impaired performance was not due to reward/punishment insensitivity but rather to “myopia for the future,” or inability to select the most advantageous option from among imagined future consequences. Damasio and colleagues conclude that these findings support the SMH.

Importantly, fMRI scans of subjects engaged in gambling tasks have generally supported Damasio’s research with skin conductance. Fukui et al. (2005) scanned the brains of 14 healthy adults engaged in the Iowa Gambling Task (IGT) in order to determine the neural correlates of risk anticipation in decision-making, where risk anticipation is the subjects’ anticipatory responses to selections from the disadvantageous decks. They found that the risk anticipation
component of the task, which was determined by subtracting data about safe decisions (i.e. selection from the advantageous decks) from data about risky (disadvantageous) decisions, activated the medial frontal gyrus of the PFC. Moreover, the subjects’ net scores in the task were significantly correlated with the magnitude of the medial PFC activity during risky decisions. Other imaging studies have also reported activation of the medial frontal cortex during risky decision-making (e.g., Akitsuki et al. 2003; Critchley et al. 2001), though some studies appear to implicate more caudal regions of the PFC (e.g., Elliott and Dolan 1998; Rogers et al. 1999).

Rogers et al., for example, reported increased regional cerebral blood flow (rCBF), in the anterior part of the middle frontal gyrus, medially in the orbital gyrus, and posteriorly in the anterior part of the inferior frontal gyrus, for a task in which subjects scored more points for predicting improbable outcomes and fewer points for predicting more probable outcomes (1999).

rCBF is a measure of the flow of blood to specific regions of interest in the brain during an activity. It is correlated with increased neural activation. These results are important both because they implicate some of the same regions of the PFC in risky decision-making as previous studies and also because the orbital and medial PFC (OMPFC) have important output pathways connecting to the medial and central areas of the ventral striatum (Haber et al. 1995). This provides an anatomical link between the frontal cortex and the striatum, which is the major input station of the basal ganglia, which is implicated in motor control and learning.

O’Doherty and colleagues (2001) reported similar results in an fMRI-based study of reward and punishment. Subjects were told to accumulate as much money as possible by choosing between two different stimuli, one rewarding and one punishing. Though money could be won and lost when selecting either stimulus, choosing the rewarding stimulus (S+) yielded
larger rewards and smaller punishments overall and choosing the punishing stimulus (S-) yielded the opposite. Through trial and error subjects determined which stimulus was more profitable (the acquisition phase), and had to learn to reverse their selections when the experimenters reversed the reward and punishment contingencies (reversal phase).

The four events of interest were (1) large rewards and (2) small punishments in the acquisition phase, and (3) large rewards and (4) small punishments in the reversal phase. The brain scans showed that the lateral area of the OFC was activated by a punishing outcome while the medial OFC was activated by a rewarding outcome. The magnitude of the neural activation reflected the magnitude of the reward and punishment received. Additionally, the medial OFC, which showed increased activation to reward, showed decreased BOLD signal to punishment while the lateral OFC, which showed increased activation in response to punishment, showed decreased BOLD signal in response to reward (O’Doherty et al. 2001: 99).

These results are important because they are consistent with the hypothesis that damage to the OFC (which here refers to the lower section of the VMPFC) results in deficient performance in emotion-related learning tasks—such as this one, and the other gambling tasks involving reward and punishment discussed in this chapter—because they impair the subject’s ability to use internal signals concerning reward and punishment to guide their choices. In other words, the function of the OFC as determined by the results of this study is consistent with the hypothesized role of the OFC in generating somatic markers, and thus appears to converge with Damasio’s research on emotion. Research on the relationship between decision-making and emotion was rapidly proliferating, and the OFC/VMPFC seemed to be the crucial link between them.
One of the most profound points of convergence between decision and emotion research in the last decade, facilitated by interest in the prefrontal cortex, involves the neural mechanisms thought to underlie both working and long-term memory as elucidated by psychopharmacology. Working memory is essentially a short-term form of declarative memory, or memory for facts and events. We have seen that Damasio’s SMH implicates the frontal cortex in the holding of integrated neural information (or signals) “online” for processing. It is now well established that the prefrontal cortex is involved in working memory. Fuster (1973) demonstrated that neurons in the prefrontal cortex of monkeys are active during the delay period in a delayed choice task in which monkeys have to remember the location of a reward in order to retrieve it after an enforced delay. Later work with human subjects and fMRI showed that both the dorsolateral PFC (DLPFC) and the ventrolateral PFC (VLPFC) are active during working memory tasks. There is even some evidence to suggest that the DLPFC is more active during tasks engaging spatial working memory while the VMPFC is more active for tasks engaging non-spatial working memory, though these functional distinctions remain controversial (Owen 1997).

These findings were part of a huge corpus of data on the neural mechanisms of memory more broadly construed. They added a new dimension to the well-established body of work on the role of the limbic system, and the amygdala and hippocampus in particular, in the production and storage of memory. Researchers have known for some time that these lower structures are

---

54 What researchers call “working memory” today was once lumped together with “short-term memory.” In 1986 Alan Baddeley, a cognitive psychologist famous for modeling memory systems, proposed to distinguish working memory from short term memory because the former emphasizes the ability of the system to manipulate stored information—to actually work with information stored on one’s cognitive “desk space”—rather than to just passively store it (Baddeley 1986; Owen et al. 1996). Consequently, most memory researchers today use “short term memory” to refer to the storage of information for short time periods, on the order of minutes, and “working memory” to refer to the short-term storage and manipulation of information.
involved in long-term memory. Some of the early evidence came from studies of patients with
damage to the hippocampus and amygdala. The most famous case is that of patient HM, who in
1953 underwent a surgical procedure for the treatment of epilepsy that resulted in the bilateral
removal of most of his medial temporal lobe (MTL), the portion of the brain containing the
hippocampus and amygdala. Following surgery, HM suffered from severe anterograde amnesia,
a condition in which one is unable to form long-term declarative (both semantic and episodic)
memories – memories for facts and events. We now know that this is because the hippocampus
and amygdala are structures crucial to the formation of long-term memory.

Where the short-term working memory associated with the PFC involves information
storage on the order of minutes, long-term memory involves storage on the order of days and
years. One of the neural mechanisms thought to underlie the formation of long-term memory is
a cellular phenomenon known as long-term potentiation (LTP). Roughly, LTP is the process
through which the transmission of chemical signals across a synaptic connection formed by two
neurons that regularly communicate with each other is enhanced. The discovery of LTP was a
watershed event in the study of learning and memory. Today, researchers are in the process of
designing and testing drugs that can both disrupt and enhance the molecular mechanisms
involved in LTP and concomitantly the formation of memory.

One such drug now being investigated is propranolol, a beta-adrenergic antagonist (“beta-
blocker” or “β-blocker”) originally used for the purpose of treating hypertension. Beta-blockers
are a class of drugs that work by antagonizing beta-receptors on cholinergic neurons. That is,
they bind to the beta-receptors without triggering a response, thereby blocking or dampening
activity at the site. More specifically, propranolol blocks the activity of epinephrine (adrenaline)
and norepinephrine (noradrenaline) on two types of the beta-receptor (β1 and β2). Because
epinephrine and norepinephrine are the hormones that underlie the “fight or flight” response of the sympathetic nervous system, beta-blockers decrease the effect of excitement on heart rate, the dilation of blood vessels, and other sympathetic functions. For this reason propranolol has been used to treat not only hypertension but also conditions that involve increased sympathetic activity such as stage fright. One of its recent experimental uses, however, has been in the treatment of post-traumatic stress disorder (PTSD), which involves dysfunction of the mechanisms responsible for encoding emotion-laden memories. And this brings us back to the amygdala. Previously I mentioned that the amygdala is the part of the limbic system closely associated with emotion. Its anatomical connections with the hippocampus and VMPFC were thought to support Damasio’s SMH. Now research with PTSD is adding additional support to one of Damasio’s key insights and the idea that we have been developing here: the importance of understanding the contribution of emotion and memory to executive function and decision-making in particular.

Amygdala dysfunction plays a central role in the development of PTSD. The amygdala is widely studied for its role in what neuroscientists call “emotional memory,” a non-declarative form of memory that mediates preferences and aversions (Squire et al. 2003). Emotional memories are generally non-conscious and independent of the declarative memories for the events in which they are rooted. In other words, emotional memory can be dissociated from both procedural and declarative memory. This has been demonstrated in triple dissociation experiments in rats (e.g., McDonald & White 1993). But it is also evident from the simple observation that we tend to remember emotionally arousing events more vividly than emotionally neutral events.
Emotionally arousing events are those that activate the sympathetic nervous system and the hypothalamic-pituitary-adrenal axis, a set of connections between the hypothalamus, pituitary gland, and adrenal glands, which release glucocorticoids (cortisol in particular) and epinephrine (i.e., adrenaline). The release of these hormones results in increased levels of norepinephrine and the familiar effects of the fight-or-flight response, such as increased heart rate and blood pressure, increased blood flow to the muscles, etc. This process is activated by signals sent from the amygdala to the hypothalamus.

Activation of this response tends results in the embedding of emotional memory, as seen for example in fear conditioning. When animals receive a mild shock following the onset of a tone, they come to associate that tone with the shock. When two different tones are used, and only one is paired with the shock, the animals come to associate the one tone with the shock and, of course, not the other. Studies have repeatedly shown that once the association has been learned, which happens relatively quickly, the animals begin to show a behavioral fear response upon hearing the shock-paired tone but not the benign tone. Importantly, following training, lesions to the amygdala eliminate the learned fear response.

The mechanisms by which these emotional experiences are embedded in memory have been extensively studied. It is now known that neurohormones are the key element in emotional learning and memory. Animals injected with norepinephrine or glucocorticoids following fear conditioning show behavioral indications of facilitated memory. Moreover, the administration of drugs (β-adrenergic antagonists) that interfere with these neurohormones has been shown to disrupt memory (McGaugh et al. 1992; Squire et al. 2003). This suggests that glucocorticoids (in particular cortisol in humans) and epinephrine both contribute to the body’s normal response to
stressful events and aid future responses to stress by enhancing declarative memory (Cahill et al. 1994). And this they do via β receptors.

PTSD involves dysfunction in the mechanisms through which emotional memories are encoded and stored. The normal responses to stress outlined above are generally helpful because increases in norepinephrine levels leads to an increase in attention and focus, and the physical changes such as increase blood flow to muscles that prepare organisms to deal with the environmental stressor. But it is equally well documented that an organism’s efficiency in responding to the stressor exhibits the inverted U-shape characteristic of the Yerkes-Dodson law in psychology (i.e., performance increases with arousal up to a point before too much arousal begins to hinder performance). In the case of PTSD, extended release of stress hormones in response to a stressful event can lead to long-term changes in brain structure and function with problematic effects on behavior. I have so far simplified the function of the HPA axis. In response to stress the hypothalamus releases corticotropin releasing factor (CRF), a hormone which in turn regulates the production of adrenocorticotropin releasing hormone (ACTH), and that in turn regulates the release of glucocorticoids (cortisol in humans). Patients with PTSD show elevated levels of CRF, indicating increased activity in the HPA axis (Bremner 2004).

One consequence of this elevated activity appears to be decreased hippocampal volume, i.e., a “shrinking” hippocampus. Based on evidence in animals that glucocorticoids decrease levels of brain-derived neurotrophic factor (BDNF), a protein involved in the growth of new cells in the hippocampus and other brain areas, researchers decided to investigate the memory functions of patients that had been diagnosed with PTSD. Significant declarative memory deficits led them to investigate hippocampal volume (Yehuda et al. 1993). MRI scans showed that the right hippocampus of PTSD patients was 8% smaller than that of normal controls.
These results have been supported by similar findings (e.g., Sapolsky et al. 1990; Bremner et al. 1997). The hippocampus as we have seen is one of the key structures implicated in memory encoding and storage, and much of the research on LTP—the molecular mechanism thought to underlie learning and memory—has been conducted on hippocampal cells.

The discovery of these mechanisms ultimately led to the idea that drugs that block the action of stress hormones at the β receptors could be used to disrupt the consolidation of emotional memories. Cahill et al. (1994) administered propranolol to subjects prior to viewing emotionally neutral and emotionally narrative stories and tested their memory for the story one week later. The stories were the same as those used previously in a study that had demonstrated an enhancing effect of emotional arousal on memory (Heuer & Reisberg 1990).

The results showed that subjects in the placebo condition recalled significantly more of the emotionally arousing story than those who had received propranolol (Cahill et al. 1994: 703). Moreover, propranolol did not impair recall for the neutral story, which indicates that the memory impairment for the emotional story cannot be explained by inattention or sedation caused by the drug. Other studies have found similar effects (Strange et al. 2003). This suggests that propranolol works by disrupting the neurohormone-dependent consolidation of emotional memory. Experimental research with the use of propranolol to disrupt the re-consolidation of traumatic emotional memory—which appears to depend upon some of the same neural mechanisms as consolidation—in patients with PTSD is now underway (Glannon 2006; Pitman et al. 2004).

At about the same time, researchers began studying propranolol for its effects on decision-making. This parallel development offers a unique look at the way in which
independent lines of molecular and cellular neuroscience research can gain support from work conducted at higher levels. In this case they are beginning to converge upon a more complete account of the role of emotion in decision-making. Disruption of the cellular and molecular mechanisms by which emotional memories are encoded appears to result in deficits of decision-making in conditions of increased arousal.

Rogers and colleagues (2004) investigated the effects of propranolol on the decision-making of healthy subjects in a gambling task while monitoring vital signs as well as changes in mood. The gambling task involved a series of trials in which the subject had to choose between two histograms, the heights of which represented the probability of winning a specified number of points (typed in green ink above the histogram) and losing those points (typed in red ink below the histogram; 2003 Figure 1 p.159). In each trial one of the two gambles served as the “control” gamble (yellow histogram), which carried a .50 probability of gaining 10 points and a .50 probability of losing 10 points. The other gamble in the trial was the “experimental gamble” in which the probabilities of winning varied between high (.66) or low (.33), and the possible gains and losses varied between large (80 points) or small (20). This yielded eight types of experimental gamble from a completely crossed design (i.e. two levels of probability were completely mixed with two sizes of gain (large and small) and two sizes of loss (large and small).

In each trial two histograms appeared and subjects pressed a key to choose between them. Moreover, the researchers included two extra types of trial in which the histograms were “losses only” and “gains only.” In losses only trials, for example, subjects were presented with a guaranteed loss of 40 points versus a .5 chance of losing 80 points and a .5 chance of losing nothing. In gains only trials, for example, they were presented with a guaranteed win of 40
points versus a .50 chance of winning 80 points and a .50 chance of losing nothing. There were ten types of trial in total presented pseudo-randomly in four blocks of trials. At the beginning of each of the four blocks the subjects were given 100 points and told to increase the amount as much as possible.

The results showed that, first, all of the subjects chose the experimental gamble more often when the probability of winning was high compared to low, and that this pattern was not altered by propranolol. In other words, when the probability of winning the gamble was high, the subjects in both the control and propranolol groups made the same decisions. Second, all of the subjects chose the experimental gamble less often when the possible losses were large compared to small. This pattern was not altered by propranolol in cases in which the probability of winning was high (Figure 2a), but it was significantly altered when the probability of winning was low (2b). That is, subjects in the propranolol group failed to choose the experimental gamble less often when the probability of winning was low and the potential losses high. This indicates that subjects taking propranolol showed attenuated discrimination between the magnitude of possible losses when the probability of winning was relatively low and the probability of losing relatively high (2004, Figure 2 p.161 esp. 2b).

These findings indicate that the beta-blocker propranolol interferes with the decision-making of healthy subjects by apparently making them less sensitive to the magnitude of a possible loss when it is more likely that they will lose than win. Put simply, the drug appears to make healthy subjects considerably less sensitive to the possibility of loss. Importantly, the mood and subjective experience measures, which were taken using standard cognitive assessment scales, indicated that the effects could not likely be attributed to generalized sedation since the changes in mood and subjective experience reported by propranolol subjects were
statistically insignificant, though of course propranolol subjects showed decreased heart rate – an expected physiological result with the use of a beta-blocker for the reasons discussed above.

The result that a beta-adrenergic antagonist, propranolol, both impairs recall for an emotionally arousing story and impairs sensitivity to the magnitude of a potential loss when that loss is more likely to occur than not suggests a common mechanism at work in these tasks. No doubt isolating that mechanism precisely will require further research, but two observations suggest an important role for emotion in both tasks, which is sufficient for my purpose here.

First, research consistently shows that gambling tasks in which subjects experience loss and reward activate the same brain regions and pathways that are implicated in emotional arousal. This has been established with both fMRI and lesion studies involving patients with damage to the VMPFC and amygdala.

Second, as we have already seen, the central mechanisms through which propranolol effects behavior at the cellular and molecular level are well understood. This means that any explanation for the present convergence of data intended to supplant the hypothesis that there is common role for emotion systems in these tasks will have to invoke the established specific activity of beta-adrenergic receptors without directly invoking the neural networks implicated in emotion. Even if it is at present too early to deny the very possibility of such an alternative explanation, given the current data it is very difficult to imagine what such a hypothesis would even look like. It would seem to require a mechanism—a neural pathway—through which beta-adrenergic antagonism affects decision-making and emotion-laden memory without directly relying upon the well-established anatomical connections between the prefrontal cortex, limbic system and somatic nervous system.
For example, an alternative hypothesis might be that propranolol blunts somatic responses without impacting the frontal networks implicated in conscious emotional feeling. But this hypothesis, or any which takes a similar line, already appears to be ruled out by the results from Cahill et al. (1994), which showed that the drug blocked the subjects’ memory for the story but did not block their initial emotional reactions to it. Thus lower-level research lends support to the hypothesis, repeatedly suggested by higher-level psychological research, that emotion systems make a substantive contribution to the mechanisms of decision.

It will be useful to briefly review the main points of the discussion in this section. I set out to place the needed link between the emerging model of decision-making in neuroeconomics and the traditional philosophical disputes about moral cognition and motivation, particularly the Humeanism dispute, that have recently received a great deal of attention in metaethics and moral psychology. The key idea is that emotion provides one crucial link. The data form human subjects reviewed in this section indicate that (1) emotion makes a crucial contribution to decision-making and thus bears directly on the research from neuroeconomics; (2) the extensive convergence of data ranging from the cognitive to the molecular is revealing details about the neural mechanisms by which that contribution is made. In the final chapter I will try to establish the link between decision and emotion research and philosophical disputes about moral judgment and motivation directly. I will argue that, because research on these mechanisms are implicating the same structures, networks and pathways in social and moral decisions, it is emotion that gives us the needed link between the neuroscience of economic decision, especially neuroeconomics, and contemporary philosophical disputes about moral judgment, decision and motivation. This massive convergence of data strongly suggests that empirically adequate theories of the nature of moral cognition, judgment and motivation must now also answer to the
neural mechanisms that underlie decision-making and motivational processes in the mammalian brain. It is time that naturalistic treatments of philosophical disputes about moral cognition begin working to develop continuity with these research projects.
6. The Science of Moral Decision

§1. Preliminaries

In the last chapter I outlined in some detail a new model of decision-making that is now emerging in the field of neuroeconomics. I then argued that the data from cellular and molecular neuroscience from which the model was initially constructed is convergent with much of the data on judgment and decision-making found in higher level sciences of the mind and brain including clinical psychology, psychopharmacology and cognitive neuroscience. I concluded by arguing that emotion makes a crucial contribution to the mechanisms of decision detailed by the model.

Recognizing the contribution of emotion is important because it lays the foundation for understanding how a model of decision in general gets connected up with moral and social decision contexts. In this chapter I complete the empirical project by connecting that model of decision to specifically moral contexts directly. If the empirical data is to be truly relevant to the Humeanism dispute, which is concerned with the relationship between moral judgment and motivation, then I need to show that the mechanisms of decision it suggests are applicable in moral contexts.\textsuperscript{55} In other words, I want to show that the different branches of neuroscience are converging toward a single, unified model of decision-making that includes moral decision. It will be to this model that empirically respectable philosophical accounts of moral agency must ultimately answer.

I conclude this chapter by returning to the questions with which we began: where in this forthcoming model of moral decision might we find the folk roles of belief and desire, and are those roles located in such a way as to vindicate the theories which have been generated by

\textsuperscript{55} Although, again, this may be even more than is really required to relate the neuroscience of decision to the neuroscience of moral decision. See the discussion at p. 204.
philosophical disputes entrenched in the FP tradition such as Humeanism and anti-Humeanism? Functionalist philosophers like Jackson, Pettit, and Smith have insisted that these folk roles must appear in future neuroscience. But I will argue that these roles are not properly located.

If I am successful, it should then be much clearer why I think the physiological constraints placed on philosophical theories by the cognitive and neural sciences need to be considered prior to formulating a philosophical framework. Philosophical disputes about the nature of moral agency and moral motivation in particular are, to borrow a phrase from Stich (1983), radically unprepared for the empirical constraints that the neurosciences are likely to place upon them. For this reason, most forms of contemporary philosophical naturalism have failed to uphold Flanagan’s principle of minimal psychological realism (Flanagan 1991). Or so I shall argue. Moral philosophy is in need of a turn toward a more genuine methodological naturalism.

§2. Some Cautionary Remarks About the Science of Moral Decision

The basic framework for a nascent model of decision is nearly in place. As a last step before considering the philosophical implications we must turn to research on decision-making in moral and social contexts specifically. For the model of decision under consideration to be relevant to the Humeanism dispute, it will have to be applicable to contexts involving norms and values. This will involve the growing body of literature on both the so-called “moral brain” and “social brain.”

The idea that the model under discussion is relevant to the Humeanism dispute will gain plausibility as the structure- or network-to-function mappings considered in the context of decision-making reemerge in the context of moral choice or decision. In other words, it will
need to be the case that the mechanisms of decision can be directly linked to the mechanisms of moral decision. Alternatively, we must be open to recognizing that the plausibility of the idea will be diminished if it turns out that the current work on moral and social decision implicates distinct regions from those previously discussed, or if no direct relationship between the mechanisms is forthcoming.

There are a few important points to keep in mind through this review of the moral decision literature. First, there is no morality center or region in the brain. This has been the conclusion of nearly every review of the neuroscience of morality literature in recent years (e.g., Casebeer 2003; Casebeer & Churchland 2003; Greene 2003), and, as we shall see, for good reason. The mechanisms of moral agency, so to speak, are no more localized than the mechanisms of decision-making, even if the constitutive component functions can be localized.

There is a related worry, which seems to take the same observation and move in the opposite direction. Moral function cannot be localized to a particular brain region not just because moral capacities require the recruitment of diffuse neural mechanisms but also because moral behavior is just not a property of brain parts. In other words, the kind of project I am pursuing here looks like it might be in danger of committing the “mereological fallacy.” This is the fallacy of ascribing to brain parts properties that can be coherently applied only to organisms (Bennett & Hacker 2003). For example, according to Bennett & Hacker, it is incoherent to claim that the part or parts of the brain whose activity is correlated with decision-making is the whole

---

56 Of course the following discussion also applies retroactively to the scientific data already discussed. I have saved this cautionary discussion for this point because of the nature of the data to follow. Though we should always be cautious and skeptical about dramatic scientific results, this is, I think, especially important in assessing data from “higher level” science, such as data from fMRI research, which is particularly prone to poor interpretation and explanation by speculative theory.
of decision-making, an activity that can be coherently attributed only to sentient, cognitive creatures.

Though my project (including much of what I have so far said) no doubt falls into the general class of projects that Hacker attacks, I hope to show that a more careful assessment of empirical data in conjunction with a turn to methodological naturalism in philosophy can help to avoid some of the confusions about which Hacker is rightly concerned. But for reasons that have already begun to emerge, and will continue to emerge, I reject his conclusion that the proper solution is to drive a wedge between science and philosophy once and for all.

As we continue to see how a more careful use of legitimate empirical data can inform philosophy (rather than vice versa) we can show that most of the conceptual confusions can be removed. The Humeanism dispute and the model of decision under discussion provide a useful example. What we must do in the Humeanism case, and what philosophers have frequently failed to do when it comes to neuroscience in general, is to distinguish between casual descriptions of cognition and behavior and the physiological mechanisms that make those activities possible. We tend to commit the mereological fallacy to the extent that we assume that neuroscientific explanations somehow exhaust or replace the levels of description for some cognitive or behavioral phenomenon. Bennett and Hacker are right to observe that this is frequently the case in the philosophical literature. For example, to claim that a brain scan or some psychological study can be used to vindicate a Kantian theory of moral agency is to commit something much like the mereological fallacy. This is because the application of Kantian norms is just not a property of a brain function, and it makes no sense to talk that way.

Still, what we can learn from scanning the brain of a subject engaged in a moral judgment is something about the mechanisms that underlie those judgments. In turn we can use that
information to determine which are the relevant causal factors in the production of a moral judgment. And these may well turn out to have absolutely nothing at all to do with any philosophical theory of moral judgment. It is a mistake to just assume that the relevant causal mechanism for some phenomenon of interest, say moral judgment, must make an appearance in the language used to describe that phenomenon at any level whatsoever. Though it isn’t quite their point, this is the sense in which I think Bennett and Hacker are right to worry about the mereological fallacy.

The solution, however, is not to drive a wedge between science and philosophy but rather to begin by determining the level or levels of description at which the relevant causal mechanisms enter the picture. In the case of the Humeanism dispute the point is that we lack any reason to accept its basic assumption, which is that the relevant causal mechanisms of moral motivation must make an appearance in the language used to describe it at the level of everyday behavioral description. If they don’t, the dispute will be incoherent. In order to know, we need to start by looking at the existing science.

Thus many philosophers draw the wrong conclusion from Bennett and Hacker’s helpful observation: what the mereological fallacy reveals is not the shortcomings of genuine philosophical naturalism but rather the shortcomings of half-hearted kind of naturalism that is ubiquitous in metaethics. The wedge—the distance between science and philosophy—is part of the problem, not a solution to it.

The point about levels of description raises one final issue before turning to the rest of the data. The diffuse cellular networks implicated in moral cognition and behavior are most accurately understood by appreciating several levels of investigation simultaneously, and in particular by paying careful attention to the too-often neglected (at least in philosophy) “lower
levels” – the levels of cellular communication, and the functions of individual cells and molecules. This is important because nearly all of the current research immediately concerned with moral cognition and behavior takes place at high levels, i.e., most of it uses fMRI or PET scan technology, which collect information about neural activation during cognitive tasks from relatively large regions. From this it is easy to draw one of several mistaken conclusions—advanced by both proponents and critics of ethical naturalism—currently popular in the philosophy, cognitive science, and economics literature.

The most popular among them is that fMRI is the “new phrenology” (e.g., Harrison 2008). Briefly, the idea is that fMRI research depends upon some of the same faulty assumptions that made phrenology a pseudoscience, most notably the modularity of mind hypothesis and the ability to make “reverse inferences” from evidence of brain activity in a particular region to the activation of a labeled cognitive process. The modularity of mind hypothesis is the familiar philosophical idea that the mind is made up of discrete modules or faculties that are located in discrete regions of the physical brain. In reverse inferences we reason backwards from evidence of brain function to the subject’s use of a particular cognitive function. For example, “if the ventromedial prefrontal cortex is active then the subject is engaged in decision-making” is a claim that infers the engagement of a cognitive function on the basis of evidence for physical activity. The inference is made on the basis of correlations between decision-making and ventromedial prefrontal cortex activity in other studies. At best these claims about which cognitive functions have been engaged on the basis of brain scans are abductive inferences (i.e. inferences to the best explanation). At worst they are instances of

57 I don’t intend this as a direct argument for reductionism. But it’s worth noting as we consider several kinds of data that the quest to understand in greater detail the nature of a wide variety of phenomena, from disease to decision making, always seems to head in the same direction.
affirming the consequent. Harrison, and many others, takes these to be serious obstacles to fMRI research.

Addressing these complaints in specific terms is a large project that I won’t pursue here. I am to some extent sympathetic to Harrison’s concerns anyway. The conclusions that scientists and philosophers draw from fMRI data are, as we will see shortly, sometimes unwarranted or overextended. Still, I want to offer one reply in more general terms for two reasons: first, I think we can make use of fMRI technology despite its limitations, and second, I think broad attacks on fMRI research rely on some of their own problematic assumptions about the nature of scientific inquiry that are prevalent in traditional (nonnaturalistic) philosophical circles.

The general reply is this: the comparison with phrenology, as I see it, fails to distinguish between empirical methods of data collection (the “science” of phrenology) and hypothesis generation and data interpretation (phrenology as a school of thought or theory). This reply invokes the distinction that I drew in Chapter 1 between science as method and science as metaphysics. There I argued that ethical naturalists have generally assumed the primacy of metaphysical naturalism at least in large part because they tend to view science as an ontology-centered form of inquiry. But science is a method, not a set of metaphysical claims, and this is why a commitment to the natural method ought always to be the mark of the ethical naturalist. Similarly, I think critics who lump all fMRI-based research programs together under the category of “pseudoscience” or the “new phrenology” generally overlook this distinction.

Phrenology was a pseudoscience not just because its assumptions were unsupported but also because it consistently failed to generate legitimate empirical data. Had it been properly conducted, it would have generated almost nothing worth reporting about the relationship between cranial topography and mental function. This is where the analogy between fMRI and
Phrenology breaks down. Phrenologists sought to localize features of the mind, specifically personality traits, to regions of the brain by measuring the features on the surface of the skull. Like fMRI, this depends on several assumptions about brain development, structure, and function. Unlike fMRI research, most of those assumptions were entirely lacking in independent empirical support. For example, phrenology presupposes that the development of topographic features of the cerebral cortex determines the topographic features of the cranial bones. For this to be true it would have to be the case that soft brain tissue presses upon and eventually changes the shape of the skull that protects it. Unsurprisingly, the mechanisms by which this was supposed to happen were never adequately explained, not to mention that their discovery certainly didn’t serve as the impetus for the development of the field.

More problematic still, phrenology seems to have assumed that most of the interesting personality features must be localizable to, or at least mediated by, the outermost layer of the cortex, the part that comes nearest to the cranial bones (and presumably does the shaping). Even the most controversial assumptions underlying fMRI research enjoy a much more reasonable degree of independent support than this. For example, even if the precise relationship between the BOLD signal of fMRI and neural activity remains under investigation, few deny that more active neurons use more energy (i.e. metabolize more nutrients) than less active neurons.

This is an important difference. In the case of phrenology, the assumptions about the mechanisms underlying the correlations have no basis in neurology or any other lower level of investigation. Soft brain tissue (which is coated by the three layers of the meninges and separated from the cranium by cerebrospinal fluid) does not press on the skull and lacks any obvious mechanism by which to change cranial topography. The BOLD signal, by contrast (no pun intended), exploits principles of hemodynamics, which account for the physical changes that
fMRI is supposed to track. Consequently, the results of fMRI scans produced by reliable methods, unlike the results of phrenology measures, are generally repeatable and can be explained in terms of an underlying mechanism.

Where fMRI becomes the new phrenology, or pseudoscience, is not in its execution but in the theorizing involved in generating hypotheses, interpreting data and drawing large-scale conclusions about the causal mechanisms of cognitive function. Unless the researchers have demonstrably botched the notorious subtraction methods or the statistical calculations that go into the production of fMRI data (which I don’t doubt happens frequently), or unless one is willing to reject what we know about hemodynamics, the results of fMRI tell us something about the brain activity of subjects engaged in particular tasks in virtue of their methodology. The problem is that what exactly they actually tell us—the metaphysical part—is often unclear for the reasons that critics like Harrison rightly observe. But this is much less a problem with the data than it is with the school of thought. The conclusion that these criticisms warrant is not that fMRI research is the new phrenology but rather that its results are generally of rather limited use in isolation.

Most fMRI research, in moral contexts or otherwise, only becomes interesting when we have data from other levels of investigation with which to make comparisons. The fact that the VMPFC “lights up” in fMRI scans of subjects engaged in gambling tasks is, by itself, almost meaningless from a philosophical perspective. But that fact coupled with the results of lesion/ablation studies, detailed anatomical considerations about networks that include the VMPFC, evidence of function from psychopharmacology, and so on, converge upon the idea that whatever mental states or cognitive functions (if there are any at all) supervene on VMPFC activity are somehow involved in that particular task. The point is not that fMRI reveals an
active VMPFC during the Iowa Gambling Task and therefore the VMPFC is the locus of gambling function, or even the locus of a mental state/function involved in gambling. It tells us only something about the hemodynamics of the VMPFC during gambling tasks. This in turn becomes useful when we know something about the behaviors that the VMPFC is involved in based upon independent lines of research conducted at different (and generally lower) levels. And these connections should be made explicit in reviews concerned with some broader, more philosophical issue so as to avoid confusions concerning both the scope of the claims and the sources of the information. fMRI research has serious limitations, and broadly speaking such data is made applicable to philosophy only by bringing in a host of background data concerning neural structures and functions.

Unfortunately, this important characteristic of fMRI research tends to be underemphasized, particularly when aimed at a broad audience. The problem helps to underscore my larger methodological complaint in this essay. Philosophers, for example, tend to report the results of fMRI scans by listing the brain region of interest (e.g., amygdala) and following it up with a generic claim about the functions that that brain region is generally thought to be involved in (e.g., emotion), but usually neglect to explain how we know that the amygdala is involved in emotional processing, where the answer extends well beyond fMRI research. Sometimes they even leave out the brain regions entirely, which is misleading given the principles underlying fMRI.

For example, here is how Jesse Prinz—a well respected naturalist for good reason—summarizes some research on the relationship between emotion and moral judgment: “They found that when subjects made moral as opposed to factual judgments areas of the brain that are associated with emotional response were active” (2006: 30). We can be charitable and take
Prinz’s point: if area A has been implicated in emotion in some other studies, and area A is active in study S, then brain areas associated with emotion are probably implicated in S. But more caution is warranted in simplifying brain scan results in this way, even when accommodating a general audience, given the kinds of objections raised by critics like Harrison.

The results of an fMRI study, by their very nature, could not yield function-to-function mappings, or even by themselves specify the relationships between two putative psychological functions. Rather they give us correlations (as critics all too often observe) between the BOLD signal as measured in some relatively small brain region of interest and particular cognitive activities. To report fMRI results (intentionally or unintentionally) without any reference to specific brain regions, simpler though it may be, glosses over this simple point and thus opens philosophical naturalism to objections from those who would equate its methods with those of phrenology. We should avoid the practice of reporting experimental results this way, which tends to generate objections that simply restate the limits of fMRI research, the modularity of mind assumption, comparisons with phrenology, and the like, which are already well documented (e.g., Nichols & Newsome 1999). Of course on the other hand, critics who claim that fMRI data tells us *nothing* by positing analogies with phrenology seem to be as confused as philosophers who claim that fMRI data alone can establish the locus of a given cognitive function. It does neither.\(^58\)

\(^{58}\) But I don’t mean to provide support for the extreme view that the prospects for localizing many cognitive and behavioral functions are poor. They are not. We have already seen how neuroeconomics has had considerable success localizing the mechanisms of choice, especially the representation of value. Whether *all* cognitive functions can be localized to discrete brain regions remains an open question, though it is already doubtful given that plenty of functions involve several extensively connected networks. Emotion is a good example of this. But the failure to localize “emotion” (as well as many other purported cognitive states and process) to a singular region probably says more about the vagueness of the concept(s) than it does the prospects for localization of function in general. And there is no doubt that many cognitive and
The most straightforward solution to this problem is to draw upon data from several levels of investigation when discussing developments in localization and to make the relevant connections between different levels of research explicit. This practice becomes increasingly important here. Even though most of the recent scientific literature concerned with moral/social decision and judgment relies on fMRI, we must not overlook the many ways in which that literature converges at different levels. With these points in mind I turn to the data.

§3. Moral Decision and the Two-Stage Model

The crucial question here is whether and how well the data on moral and social decision converges with the data adduced in support of the two-stage model of general decision-making. Here I present some data and then offer an argument to the effect that the convergence is at least plausible and perhaps even reasonably well supported at this early stage.

The primary piece of evidence for this view is the substantial overlap between brain regions of interest in both the neuroeconomics and moral/social judgment literature. The extensive convergence between some of the lines of research in these fields is becoming increasingly difficult to ignore, particularly for philosophers who claim that their armchair theories of moral cognition and behavior developed from an analysis of folk platitudes must be relevant to or preserved by future scientific work. As we continue to uncover the mechanisms of moral agency it becomes increasingly difficult to deny that philosophical theories like Humeanism and anti-Humeanism are in fact in competition with explanatory hypotheses under behavioral functions do in fact depend on the physical features of the particular cells in a region of the brain. The cells in the retina (which is, of course, a part of the nervous system given its developmental origin in the optic nerve and hence the brain) are a particularly good example. So any objection to fMRI research that hinges on a general claim about the failure of localization is absurd. There may yet be plenty sophisticated reasons to be critical of fMRI research, but a pervasive failure of localization is not one of them.
development in the neural sciences. Unless the relevant causal mechanisms emerging from the scientific research actually appear in the philosophical theories whose primary mechanisms include FP intentional states, one or the other will have to be rejected as a relevant explanatory hypothesis. I turn now to the question whether the mechanisms being uncovered in neuroeconomics might be applicable to moral agency.

Moll and colleagues (2001) scanned the brains of subjects presented with simple claims, some of which contained moral content requiring judgment about rights, responsibilities or values (e.g., “Old people are useless” and “They hung an innocent”) and some with non-moral content involving judgments of fact (e.g., “Stones are made of water”). Subjects were asked to silently judge whether the claims were right or wrong while in the scanner. The results showed that judgments about moral claims produced increased activity in the following regions: the frontopolar cortex, right cerebellum, the right temporal pole, left OFC, left precuneus, posterior globus pallidus, superior temporal sulcus (STS) and medial frontal gyrus.

In a later study (2002), these same researchers compared regions of activation during judgments of moral claims to activation for judgments of unpleasant nonmoral claims with social content in order to determine whether the neural systems that mediate emotionally charged moral and nonmoral judgments were dissociable. Because both types of claims contained emotional content, they could determine whether any brain regions showed increased activation particularly for moral or social content. They reported that moral judgments associated with unpleasant

59 After scanning the subjects were presented with the list of claims a second time and instructed to recall their impressions during the task and to rate on a 4-point Likert scale whether each sentence was right or wrong, the amount of moral content each contained, the degree of difficulty in responding, and the emotional valence (positive or negative) of the claim, thus providing a measure against which to interpret the brain scans.
60 A gyrus is a ridge formed by bunched up cortex. Its “opposite” (so to speak) is a sulcus, which is a depressed area around a ridge or gyrus.
emotion activated the anterior part of the medial OFC while nonmoral social judgments associated with unpleasant emotions activated the lateral OFC and the amygdala (2002 Figure 1, p. 700). This of course is not evidence that the anterior portion of the medial OFC is the morality center of the brain. It does, however, suggest that that part of the medial OFC is activated in emotion-laden moral contexts. And that might be important as researchers continue to explore the various tasks in which the medial OFC is especially active, and as we learn more about the properties of the cells in that region.

Greene and Haidt (2001) have contributed to that project. They scanned the brains of subjects responding to a series of complex narratives containing putatively personal and impersonal moral dilemmas. These included “personal” (direct) and “impersonal” (less direct or perhaps indirect) variations on the famous trolley problem in which subjects must decide whether to divert the course of a runaway trolley, intentionally killing one person in order to save five. Narratives involving personal moral dilemmas contained situations in which the individual would have to directly inflict harm on another, such as by pushing one person in front of the runaway trolley to stop it from killing five people tied to the tracks. Impersonal moral dilemmas were those involving less direct harm, such as flipping a switch to divert the trolley instead of pushing a man onto the tracks.

They found that responding to personal moral dilemmas increased activation in the medial frontal gyrus, posterior cingulate gyrus, and bilateral superior temporal sulcus (STS). Impersonal moral dilemmas, much like non-moral dilemmas (a control condition), increased activation in the DLPFC and parietal cortex (BA 7/40), areas generally not implicated in emotional function but rather, as we saw earlier, in working memory and executive cognitive control.
Concerning the impersonal condition, DLPFC activation is interesting because it seems to suggest that impersonal moral dilemmas, which involve indirect forms of intervention such as flipping a switch, could actually be more like practical judgments than moral judgments. Recall that the DLPFC is consistently activated in tasks involving “executive” cognitive functions (such as testing hypotheses or identifying rules for sorting or reward distribution, as in card sorting tasks) and is widely thought to be involved in the contribution of working memory to those functions (Elliott & Dolan 1998; Goldberg et al. 1996). Greene et al. argue that it is at least plausible that indirect nature of the intervention in the impersonal condition has the effect of making the dilemma an exercise in practical reasoning. This would account for the DLPFC activation. I will return to the practical reasoning hypothesis shortly.

These results suggest a direct link between the concept of judgment and the mechanisms of decision-making in the following ways. First, these patterns of activation lend support to the idea that judgment is best construed as a species of decision-making since the subjects in the study faced the practical judgment about what it was best to do in the context of the trolley problem. Note the similarities with the operational definition with which we began – moral judgment is a species of decision, decision about what it is best, or perhaps optimal, to do given the circumstances.

Second, the activation of the parietal cortex is by now also familiar, and serves as another example of data convergence. Impersonal and non-moral dilemmas activated the region of cortex discussed above as involved in fronto-parietal networks essential to the execution of choice selection. BA 7 is situated just behind the primary somatosensory cortex and as we saw earlier is the part of the parietal cortex involved in bringing together visual and motor information (visuo-motor coordination).
And in a third important point of convergence, the DLPFC plays an important role in decision processes as described above. The DLPFC (BA 9/46) is anatomically connected with areas of cortex involved in the processing of sensory, motor, and affective signals. DLPFC neurons have been shown to modulate their activity in response to expectations about rewards (Leon & Shadlen 1999). For example, neurons in the DLPFC will respond more when a monkey expects a larger reward in a spatial working memory task (Kobayashi et al. 2002). And Watanabe (1996) showed that neurons in the DLPFC are responsive to both a particular award and to the way in which the award is given. In other words, these neurons encode the outcome of goal-directed behavior.

The scans of subjects in the personal condition support the involvement of emotion in moral judgment as follows. The medial frontal gyrus encompasses Brodmann’s areas (BA) 9 and 10. This is roughly the same region previously discussed as the frontopolar cortex, very near the VMPFC. Moll et al. (2001) reported that the area was active as subjects thought about simple moral claims. The posterior cingulate gyrus is a region of cortex corresponding roughly to BA 31, near the retrosplenial region of the cingulate cortex at BA 30, while the STS is a long sulcus in the temporal lobe whose caudal tip corresponds to BA 39. Notably, the STS has reciprocal connections with the amygdala, which is itself reciprocally connected with the OFC. The posterior cingulate and retrosplenial cortex are very frequently reported active in tasks involving emotion. In fact the retrosplenial cortex is the region most consistently activated by emotionally salient stimuli (Maddock 1999). The STS and inferior parietal region (BA 39), like the frontopolar cortex, were active as subjects thought about simple moral claims (Moll et al. 2001).
The STS and inferior parietal region have also become the focus of research on social cognition. The results have so far suggested that the area is involved in the representation of socially significant information, especially social information acquired from the perception of motion, such as the intentions of others. These abilities constitute one part of “theory of mind” (ToM), a general term for the ability to perceive and process and ultimately represent information about the mental states of others, including their dispositions and intentions. The STS is known to be involved in the direction of gaze. Monkeys with lesions to that area exhibit deficits in the ability to perceive the direction of a gaze (Allison et al. 2000; Campbell et al. 1990). Deficits of this type are sometimes seen in humans with prosopagnosia, a disorder characterized by difficulty recognizing or identifying faces (Perrett et al. 1988). Moreover, single-cell recording studies have shown that a select subset of cells in the STS fire vigorously in response to certain kinds of head and face movement and not others. Hasselmo et al. (1989) reported a vigorous response of a single cell in the lower bank of the STS to ventral flexion (downward movement toward the chest) of the head, downward movement of the eyelids, though not to direction of gaze. The researchers hypothesize that these downward movements are important because they are involved in the breaking of contact between monkeys engaged in dominance interactions (Allison et al. 2000; Hasselmo et al. 1989).

Additionally, damage to the temporal lobe, across which the STS extends, is a possible cause of not only prosopagnosia but also Capgras syndrome, a disorder characterized by the delusion that a familiar person, usually a loved one, has been replaced by an imposter of identical appearance. Capgras syndrome is sometimes said to be the reverse of prosopagnosia because Capgras patients recognize faces though they deny identity, while prosopagnosics fail to recognize faces but often can identify others through alternative sensory modalities, such as by
voice recognition (see especially Ellis & Young 1990). There is a good deal of evidence that the inferior region of the temporal lobe is the locus of a group of cells that are specialized to respond to faces, a fact which may help to explain the symptoms of these disorders.

The inferior temporal (or inferotemporal; IT) cortex is generally implicated in the recognition of complex forms. This has been established through lesion studies, both with humans in clinical studies and monkeys in experimental research. Removal of the IT cortex in monkeys impairs shape and pattern recognition without affecting other basic visual functions. Moreover the cells in this area have some features that make them particularly suited for face recognition. For example, nearly all of the cells include in their receptive field\(^6\) the foveal region, the region in which fine discriminations are made in the eye. Foveal vision is crucial for any activity requiring perception of sharp detail, such as reading. Additionally, unlike the cells in both V1 (also called the striate cortex) and extrastriate visual areas, the cells of the inferior temporal area do not exhibit retinotopic organization. That is, they do not form a map of the visual field through the arrangement of receptive fields like the cells in V1, the extrastriate visual areas, and the superior colliculus. They also have very large receptive fields, sometimes large enough to include the entire visual field. These features make it possible to recognize the same detail or feature anywhere in the visual field. Hence they likely account for our ability to recognize the same object, such as a face, regardless of its location in space – an ability known as position invariance (Kandel et al. 2000).

---

\(^6\) The concept of a receptive field (RF) is probably among the more abstruse concepts in neuroscience. For our purposes it is sufficient to define an RF generally as just the portion of space in which the presence of a stimulus affects the firing of a neuron. Depending upon the sensory modality, that portion of space could be a part of the visual field, an audio frequency, or a location on the skin.
More directly, some cells in the IT cortex respond selectively to particular stimuli, including hands and faces. Kobatake and Tanaka (1994) recorded from cells in the IT cortex of macaque monkeys as they were shown stimuli that resembled the face of a monkey to varying degrees. By gradually changing the features of the stimuli, researchers were able to determine the features of the face to which the cells were responding. In other words, they were able to determine which features were indispensable to cellular activity, and therefore to the recognition of a face. Unsurprisingly, an encircled bar and dots were crucial to neural activity (Kobatake & Tanaka 1994 Figure 4B, p.859).

The existence of cells in the IT cortex selective for faces, coupled with clinical work with prosopagnosics and patients with Capgras syndrome suggest that the IT cortex is responsible for, or at the least a crucial contributor to, face recognition. Even so, the precise causes of prosopagnosia and Capgras syndrome remain under investigation. The lesions present in prosopagnosics consistently extend from the occipital lobe to the inner temporal lobe bilaterally.

Less is known about the precise kind of damage implicated in the Capgras delusion. However, one of the leading hypotheses suggests that the disorder is caused by a disruption of communication between the temporal cortex and the limbic system (Hirstein & Ramachandran 1997; Ramachandran & Blakeslee 1998). Prosopagnosics exhibit autonomic responses (measured by skin conductance responses or SCRs) to familiar faces though they cannot consciously recognize those faces (Ellis et al. 1997). Patients with Capgras syndrome, by contrast, consciously recognize familiar faces but seem to lack the appropriate emotional or autonomic response. This suggests that the Capgras delusion may be caused by the failure to generate an emotional response to the processing of familiar stimuli. In the absence of the
emotional response, the subjects become convinced that the familiar person must in fact be unfamiliar.

If this is correct, we seem to have additional support for the now familiar dissociation between the conscious processing of sensory information and the unconscious processing of emotional information, a result consistent with Damasio’s work. In any case these results concerning facial recognition seem to support Greene’s claim, namely that the STS and inferotemporal cortex are directly involved in, or perhaps even the locus of, the representation of personhood. If so, it is possible that the personal moral condition in that study activates regions involved in the representation of personhood while the impersonal condition does not.

Many of the foregoing considerations led Greene and colleagues (e.g., Greene et al. 2004) to put forward what is sometimes called the “dual-process view” (or “dual-systems model”) of moral judgment. On this view both cognitive and emotional processes play a crucial role in moral judgment and sometimes those processes are in competition with one another. The cognitive processes in question are the reasoning processes of so-called “higher cognition” associated with problem solving, such as in the Wisconsin card sorting task. The emotional processes are those discussed previously as emotional “feeling,” known to involve the VMPFC.

The results of their work with the personal and impersonal variations on the trolley problem as discussed above opened up new avenues of research with utilitarian and non-utilitarian moral judgments. Because it appears that the VMPFC—the frontal area that is consistently implicated in conscious emotional feeling in independent lines of research—is active for personal but not impersonal moral judgments, and thus that emotional process are involved in personal moral judgments, it seems plausible that these emotional processes are also
involved in any judgment which draws attention to considerations other than just raw mathematical calculations of consequences.

In other words, perhaps these emotional processes are involved in judgments that require something other than strictly utilitarian consequentialist reasoning. In the original version of the famous trolley problem in which subjects are asked whether they would flip a switch to divert the trolley so that it kills one person rather than five, it appears that most subjects employ basic utilitarian reasoning to conclude that one death is preferable to five even though that outcome requires active (albeit indirect or “impersonal”) intervention. When that intervention is made more direct (or “personal”), however, subjects appear to reject the conclusion of the basic utility calculation in favor of the intuition that it is somehow more wrong to push a large man off of a footbridge in order to stop the trolley – a “non-utilitarian” judgment.

Greene and colleagues hypothesized that this is likely because the personal nature of the intervention arouses emotion systems that disrupt commitment to the utility calculation. To test this, they presented subjects with dilemmas like the “crying baby dilemma”:

Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside, you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth, his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others, you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople? (2004: 390).

Notice that in either case the baby dies. The utilitarian calculation seems to suggest that smothering the child is the thing to do. If the baby cries then everyone dies, baby included. So it seems that by smothering the child you sacrifice one life to save several. Still, that one life is the life of your innocent baby, and this is supposedly where the emotion-laden reasoning process
associated with visceral reactions to personal moral violations comes into conflict with the straightforward utilitarian calculation. Judgments that result in rejecting the suggestion of the utilitarian calculation are “non-utilitarian.” Notably, the researchers suggest that in fact the subject should see that there is “nothing to lose (relative to the alternative) and much to gain from carrying out this horrific act” (2004: 390). It seems they take this to be the “right,” or rational, answer.

Resolving the conflict between the utilitarian and non-utilitarian judgments generated by the dilemma is thought to require “cognitive control,” i.e., the ability to guide thought and attention in accordance with goals and behaviors. This is sometimes thought to be the same ability required by the famous Stroop test, in which subjects are asked to name the color of the ink used to spell the name of a different color (e.g., “red” written in green ink). In both, subjects must consciously guide their attention to the relevant features and overcome obstacles to arrive at an answer. Unsurprisingly, both the crying baby dilemma and the Stroop test significantly increase the RTs of subjects compared to control trials. The Stroop task has been shown to activate the ACC in independent fMRI research, which suggests that it plays a role in cognitive control (Botvinick et al. 2001; Greene et al. 2004).

Given the patterns of activation in the previous study (Greene et al. 2001), the researchers hypothesized that the fMRI scans of subjects who ultimately chose “appropriate” would show greater activation in the areas associated with abstract reasoning, particularly the DLPFC, as compared with those who ultimately reject the conclusion of the utilitarian calculation. Given the need for cognitive control, they also predicted that the scans would show increased activation in the anterior cingulate cortex (ACC). Interestingly, the results seem to support both hypotheses.
First, difficult moral dilemmas such as the crying baby dilemma, compared to easy moral dilemmas, increased activation in the anterior DLPFC and the inferior parietal lobes (BA 40/39) and in the ACC and posterior cingulate cortex (BA 23/31).

Second, the results show increased activity bilaterally in the anterior DLPFC (BA 10), in the right inferior parietal lobe (BA 40), and in the anterior portion of the posterior cingulate (BA 23/31) in subjects who ultimately made the utilitarian judgment (2004 Table 1 p. 392). The authors note that one of the most interesting features of these results is the increased activity in the DLPFC for utilitarian judgments given that the results of the previous study showed decreased DLPFC activity in personal moral judgments.

As is nearly always the case with fMRI-based research, this data raises more questions than it answers. Most important among them: how certain can we be that these results, even if the data analysis was flawlessly executed, actually support the dual-systems view of moral judgment; and what do these results actually tell us about decision-making processes in social and moral contexts?

The first question reasserts the methodological concerns of the previous section. The answer, to repeat, is that it is far from certain, though the claims can gain plausibility to the extent that they converge with other lines of research. This remains an open question. There are several assumptions at work in the conclusions that Greene and colleagues draw from their data that seem to require additional support, including: that subjects do in fact use utilitarian calculations when they assess dilemmas; that utilitarian judgments are not just as bound up with emotional processing as non-utilitarian judgments; that subjects are genuinely able to place themselves “inside” (so to speak) the dilemma such that the emotional reaction to smothering one’s own child is comparable to a real scenario; that “cognitive control” is really at work in
these cases, and so on. Of course, establishing each of these will be nearly as difficult and controversial as establishing the dual-systems view itself. At best, then, we can try to make some preliminary assessments.

Given the data canvassed in this chapter alone, it is plausible that some foundational elements of the dual-systems view are right, though the dual-systems model itself is probably far too simplistic. The general idea that different neural systems and pathways can conflict, to focus on just one of those elements, is well supported. We have seen this not only in the Stroop case, but also in cases of rule change in card sorting tasks and in the Capgras delusion.

Still, it is not clear that the kind of cognitive control required for arbitrating between options in the crying baby dilemma is the same as the kind of cognitive control required for the Stroop test, particularly because the latter has clearly correct answers while the former does not. In my own experience taking the Stroop test, after several trials you begin to establish techniques for reducing the amount of time required to determine the correct answer, such as concentrating your gaze on individual letters and trying very hard not to read the words. You do this precisely because you already know that there is in fact a single correct answer and your goal is to find the most efficient way to reach it.

In the crying baby dilemma by contrast it is not even clear what the “correct” answer is—if there is one at all—after several hours of reflection. So while the data suggest a genuine case of cognitive conflict, it is still too quick to conclude that it is the same kind of conflict present in other tasks such as the Stroop task, and that it suggests two competing neural systems.

This last point is supported by the existence of a competing explanatory hypothesis. It is

---

62 This is of course appears to be just another manifestation of the is/ought gap. Brain scans might help us understand how individuals make moral decisions, but they can’t tell us which decision is morally “right.”
perhaps equally plausible that the conflict generated by the narrative is not the conflict between two competing neural systems but rather conflict within the processes constitutive of one system: perhaps a single system of valuation in which the subject comes to assign subjective values to the options before choosing between them. If so, there may yet be a role for emotion in the judgment, but that role may be rather different from competing with a unified system for abstract thought processes to reach a decision. It may be something more like a role in determining the subjective values of the options, or additional mechanism by which to ultimately tip the scales, as Glimcher has suggested.

After all, contrary to Greene’s assumption, it isn’t perfectly obvious that—at least in an actual rather than hypothetical case—there is just one clear utilitarian answer to the dilemma, which is to smother the baby. One of the standard criticisms of utilitarianism is that it is entirely unclear how to objectively assign utility points in order to carry out the calculations (e.g., Williams 1973).

Moreover, it seems to me that the crying baby dilemma, or any of the variations the trolley problem, invariably gives rise to questions about possible alternatives in real rather than hypothetical scenarios: isn’t there some alternative way to keep the baby quiet that doesn’t result in suffocation? In experimental settings, of course, the subjects are told explicitly to choose between their two options. This artifact of using hypothetical dilemmas makes it easier to test the utilitarian/non-utilitarian hypothesis, and it also makes it easier to see how the putative utilitarian calculation might be preferable. But even when we are explicitly instructed to see it this way, it is probably hard for most of us to avoid reflecting, even if only momentarily, on what else we might do if we were really in such a situation. If this natural tendency tells us anything, it probably tells us about the limits of the experimental design. The artificial nature of being
forced to choose between two unappealing options in a hypothetical case only serves to emphasize the complexity of real-life decision-making. So it is at least plausible that the complexity of decision-making, especially in moral contexts, may actually be better explained by the two-stage model that emerges from general decision research than by directly linking traditional moral theories to putative independent, competing neural pathways.

In other words, some of the explanatory speculation from Greene and colleagues may just be another instance of the quasi-naturalistic tendency to preserve the unchecked empirical assumptions built into traditional philosophical disputes. The idea that there are two basic reasoning systems that come into conflict in such cases seems ad hoc.\(^{63}\) It suits the dilemma data nicely, but perhaps lacks applicability to more general kinds of decision-making. Despite that, the data itself need not be called into question just because the framework used to explain it lacks support.

The two-stage model of decision under consideration, for instance, may ultimately help to

\(^{63}\) Interestingly, this is not the only place that Greene and colleagues have proposed two dueling systems in order to explain our failures of reasoning. In Greene & Cohen’s 2004 article about the implications of recent neuroscience research for our legal system, they conclude by proposing that the problem of free will may best be explained by an internal conflict between our “folk physics” and “folk psychology” cognitive systems. There too the proposed explanation seems ad hoc.

Moreover, proposed conflicts of a similar sort between exactly two independent systems are quite popular in the philosophical literature. For example, many philosophers and psychologists are currently engaged in a series of disputes about rationality (the “rationality wars”). The dispute concerns how best to explain the human tendency to reason poorly, a tendency exhibited by performance in the Wasson card selection task, or when dealing with probabilities. Some of these philosophers have appealed to data concerning the interplay between two systems, namely a fast, intuitive reasoning system and a slower, reflective reasoning system (e.g., Carruthers 2009; Kahneman & Tversky).

Of course, the tendency to propose two-system models isn’t in itself any real problem, though I think it suggests a simplistic treatment of neuroscience data. By contrast, one of the advantages of a two-stage model of decision is that the division of stages is more epistemic than ontological. The relevant observation is only that the mechanisms of valuation take place prior to choice selection, however we choose to divide the stages. On the other hand, dual systems models lose credibility immediately if it turns out that there are more or less than two systems.
explain the process by which we weigh all possible options, the difficulty we have in selecting among them (and hence the measurable increase in RT), and perhaps even how emotion systems can aid or influence that process—all indicated by the Greene et al. data—without forcing connections between neural systems (e.g., emotion and reason) and traditional moral theories (e.g., utilitarianism and deontology), which depends on controversial assumptions. We can and should take the data seriously. But we should be careful about forcing it to square with traditional philosophical frameworks.

And this brings me to the second question raised by this research, which is my primary concern here: the accuracy of the dual-systems view aside, what do these results tell us about decision-processes in moral and social contexts?

However we evaluate the speculative explanation for the data, Greene and colleagues seem to have uncovered important evidence for some key components of a neural model of *moral* decision: brain areas activated under conditions of cognitive conflict are active in difficult dilemmas (regardless of whether the conflict is between emotion and reason systems or between the neural representations of each of the options that rely on both reasoning and emotion processes); brain areas involved in abstract reasoning are also active in certain kinds of moral dilemmas, or are perhaps responsive to certain features of moral dilemmas; and perhaps most importantly, brain areas implicated in emotional processing appear to be activated by personal moral dilemmas. And all of these points appear to square with the data considered earlier in this chapter.

Whatever the precise mechanisms at work in subjects faced with moral dilemmas, then, it seems clear that the results at least suggest that the mechanisms of decision-making so far discussed could potentially be involved in moral and social decision. A few more empirical
considerations worth mentioning appear to support this conclusion.

Koenigs et al. (2007) conducted experiments with patients who had bilateral lesions to the VMPFC in an effort to determine whether the emotional processing contributed by that region is necessary for moral judgment. They presented the patients with a series of moral scenarios, some of which involved relatively unemotional impersonal harms such as lying on a resume, and some with highly emotional personal harms, such as the crying baby dilemma. A subset of the emotional, personal harm scenarios involved conflicts between an emotionally aversive harm (e.g., smothering your baby) and the greater good (e.g., saving the townspeople from the enemy soldiers).

The results showed that VMPFC patients responded normally, i.e., their responses were in agreement with the controls, to the relatively emotionless impersonal moral harms. But they were significantly more likely to endorse a course of action involving an emotionally aversive personal harm for the greater good than the control subjects (2007: 910). In other words, patients with VMPFC damage were more willing to accept the “utilitarian” conclusion.

These findings have been replicated (e.g., Ciaramelli et al. 2007). And similar studies have been carried out with patients suffering from frontotemporal dementia (FTD), a disease characterized by deterioration of the frontal lobe and in many cases extending to the temporal lobe. Mendez et al. (2005) presented FTD patients with the two standard variations on the trolley problem (i.e., flipping a switch to kill one person and save five, and pushing a large man off of a footbridge onto the tracks to save five). The results indicated that most of the FTD patients approved of both actions while normal subjects approved only of flipping the switch (i.e., they approve of the “impersonal” harm but not the “personal” harm). These results are consistent with those from both Greene’s and Koenigs’ studies, suggesting a crucial contribution.
of the VMPFC and—at least to the extent that the VMPFC underlies emotional contributions to reasoning—emotion to moral judgment.

§4. Some Preliminary Conclusions

At the center of the preceding review is the two-stage model of decision now emerging in neuroeconomics. What I take this review to show is both that the model gains a great deal of plausibility from the convergence of several independent lines of research and that it therefore has serious implications for our ability to make sense of traditional philosophical questions about the nature of moral agency. As we have seen, the crucial link between the neuroscience of economic decision and philosophical disputes about moral judgment and motivation, and the Humeanism dispute in particular, is the extensive neural literature on the contribution of emotion to decision-making processes. That contribution is present not just in economic decision but also in decisions about social and moral norms. Moreover, the neural networks and pathways implicated in decision processes in each case are strikingly similar. Too many philosophers who claim that their theories of moral agency are empirically respectable or psychologically realistic have ignored these points. In part this is because the direct connection between the neuroscience of decision-making and the psychology of moral judgment and motivation have not been widely recognized. But this trend, I have argued, is perhaps best explained by the pervasive tendency to privilege traditional philosophical frameworks containing controversial empirical assumptions and thus to disrupt continuity between philosophical and scientific methodology.

On the one hand, many philosophers have constructed detailed theories of moral motivation and moral judgment without doing much of anything to track down the results of actual empirical research first. The Canberra Planners, as we have seen, claim that the empirical
work is unnecessary. And many naturalists have been content to develop their psychology from the armchair. Doris & Stich (2005) and Darwall, Gibbard, and Railton (1997) have lamented this fact in their recent discussions of naturalism. There is an important difference between commonsense observations that have a high degree of prima facie empirical plausibility and observations that have a high degree of empirical support. The former make for good empirical hypotheses, but they are not sufficient for constructing an empirical theory of moral agency, much less for serving as pillars of metaethical inquiry in general as in Smith’s moral problem.

On the other hand, those who have taken an interest in the actual data have generally focused exclusively on one single line of investigation. Naturalists like Kennett, or Greene, for example, tend to take a serious interest in cognitive neuroscience and fMRI data. Unfortunately though they have done too little to determine whether independent lines of empirical research, particularly from other (lower) levels of neuroscience, do more to explain their results than do armchair philosophical theories. Despite the enormous body of literature on the neural mechanisms of decision, they have been content to explain the results of fMRI scans in terms of (among other things) Kantianism and “dual-processing” systems, which seem a bit simplistic given the state of current emotion research. In fact though the most interesting feature of all of this data is the way in which it converges toward a rather unified theory of the mechanisms of decision. It would be a mistake to write these results off as a coincidence, or to ignore them in favor of a single-minded interest in the higher-level neurosciences or theories couched in the language of ordinary, everyday behavioral description.

Consequently I think that it is this two-stage model that turns out to be the best place to look for the theoretical realizers that we will need to vindicate the Humean dispute about moral motivation, and other related philosophical disputes about the nature of moral agency. We need
not wait for some future neuroscience because it is clear that much of the next decade worth of work will be devoted to filling in some of the details of that model.

To take one especially germane example, new lines of work are already underway to clarify how exactly emotion influences valuation and choice selection within the two-stage model itself. Much of this (for reasons previously discussed) involves cashing out the vague concept of emotion. Here is how the leading emotion researcher Elizabeth Phelps (2006; 2009) characterizes this new direction in neuroscience and neuroeconomics:

To date, most neuroeconomics studies have depicted emotion as a single, unified variable that may drive choice, often in contrast to a reasoned analysis of the options and their relative values (see, for example, Cohen 2005). This dual-system approach, although intuitively appealing, fails to consider the complexity of emotion or to capture the range of possible roles for emotion and affect variables in decision making. Adopting a more nuanced understanding of emotion will help clarify its impact on economic decision making and provide a basis for further understanding the complex interactions between emotion, value, and choice (2009: 234).

Metaethics and moral psychology should take note of this for at least two reasons. First, it suggests that the model is here to stay, which means that philosophers interested in engaging with the “future” neural and cognitive sciences will likely have to address it.

Second, emotion is currently a central concept in philosophical disputes over theories of moral agency. As the empirical literature continues to provide more careful and fine-grained analyses of the concept, it will likely become increasingly difficult to even begin to assess the empirical adequacy of such theories anyway. If Phelps is right, we might expect that in time the moral psychology literature will be forced to become more continuous with the sciences in order to avoid empirical irrelevance. At some point we will be forced to choose between revamping our philosophical concept of emotion in light of scientific advances—that is, to trade our philosophical concept of emotion simpliciter for more nuanced accounts of the particular structures and pathways involved like those we have just seen in the previous chapter—or else
continue formulating theories of moral agency using a concept which is either too vague to be of use or specific enough to plainly fail to capture the state of the science.

As the science of emotion progresses the possibility that we will be faced with a division between the vocabulary of commonsense and empirical psychology starts to loom large for proponents of armchair psychology. How are we to make sense of sentimentalism, Humeanism, projectivism and the like if the concept of emotion (simpliciter) undergoes substantive refinement in the next decade? And why should we be so convinced that Phelps’ point couldn’t hold just as well for beliefs and desires?

Of course, the armchair theorists will probably be content to cross that bridge if and when we come to it. So we are now ready to look at how beliefs and desires fare given the two stage model from present neuroscience. I will try to explain why the idea that neuroscience will or must ultimately preserve the structure and relevance of beliefs and desires in the best available causal explanations of moral judgment and motivation ought to strike us as outrageously presumptuous. Given the structure of the best available causal explanations of decision and motivation, we must recognize that that assumption is far more audacious and extravagant than its denial.

§5. In Search of Theoretical Realizers

The data I have reviewed over the last two chapters is poised to raise some pressing challenges to philosophers interested in vindicating traditional philosophical theories of moral motivation. These philosophers must either finally put paid to the task of locating traditional folk psychological (FP) concepts like belief and desire in our going scientific explanations for motivational processes (call this the “location project”), or they must content themselves with
formulating theories in commonsense terms (i.e., speaking from *within* the perspective of moral agency rather than *about* it, cf. Blackburn 1998) and jettison appeals to scientific data and claims about empirical respectability. But philosophers who appeal to the sciences to support philosophical theories of moral motivation have already conceded this latter project. They now owe some plausible account of how FP states might plausibly “supervene” on the neurophysiology, and more specifically on the neural mechanisms so far described.

In this section I argue against the likelihood of providing such an account along the following lines. First, the successful account of location will have to meet (at least) two obvious conditions: it must preserve both the causal relevance of FP states and the characteristics of those states which make them commonsensical and hence a mainstay in folk vocabulary. The trouble is that given the foregoing data we may have reason to expect that these conditions will tend to work against one another. Achieving one will tend to come at the cost of the other. Thus I think each of these desiderata poses a special challenge for FP realism in light of the neuroscience of decision and motivation.

It is striking how infrequently the words “belief” and “desire” actually appear in the body of literature just reviewed. Still, perhaps we can make a plausible case for the presence of beliefs and desires in the research so far presented with the help of some philosophical interpretation. I will begin with the first stage of the two-stage model: valuation.

The most straightforward way to invoke the folk roles in this stage of theory is to say that valuation is a neural (or perhaps cognitive) process whereby an agent comes to associate a given reward or option with both a degree of desirability and a belief or set of beliefs about which items or actions carry which values and the probability of satisfying her desire given her previous experience with an option, and so on. I suspect that the moral functionalist has
something like this in mind. All the low-level neuroscience in the world will do nothing but illuminate the physical processes whereby an agent comes to believe that an action or item has such and such a value and the processes whereby she comes to desire the outcomes of those actions.

We need to represent this idea more carefully by translating it directly into the language of the valuation stage of the decision model. Recall that the two stages in the two-stage mechanism for decision-making emerging in neuroeconomics are valuation and choice. In valuation, subjects assign subjective values (SVs) to individual goods or actions in a range of options. At the behavioral level, SVs can be understood as economic values calculated by quantifying the subject’s choices relative to the alternatives. At the neural level, it turns out that these SVs can be defined as the mean firing rates in action potentials per second of specific populations of neurons. These neural SVs are learned, represented, stored, and ultimately used to guide motor system activity. In the second stage, choice, this neural information concerning the most highly valued item or action is then implemented into motor pathways to guide physical action.

The neural process is thus very much like the process postulated by behavioral economists whose traditional models of economic choice explain decision making ‘as if’ it involves choosing a highly-valued option from among an array of options represented in common currency (the ‘common currency hypothesis’). Neurophysiological data now indicates that decision-making at the neural level does indeed seem to utilize a common currency. That currency is SV, which we defined previously as:

\[
(SV) \quad SV = \text{df the responses—particularly the mean firing rates—of specific populations of cells, quantifiable in real numbers whose units are action potentials per second, to each of the items or actions available.}
\]
The neural pathways and regions implicated in valuation are the ventromedial prefrontal cortex (VMPFC) and striatum, while those implicated in choice are the lateral prefrontal and parietal cortex (Kable & Glimcher 2009; henceforth ‘K&G 2009’). Recordings from cells in the VMPFC have contributed to the localization of valuation. Researchers have identified three types of neurons that respond to the values of individual goods on offer regardless of whether they are chosen (offer value neurons), to the values of goods and actions actually chosen (chosen value neurons), and to the chosen action itself (choice neurons; Padoa-Schioppa & Assad 2006, 2008). Similarly, these three types of neurons have been found in the caudate and putamen of the striatum where research indicates they track the values of actions (rather than goods; Samejima et al. 2005).

Recall that these studies also showed that the neuronal value representations were menu-invariant. That is, the neural responses are representations of the economic value of individual goods/items rather than representations of relative value, or the value of a good relative to its paired alternative. This is important because behavioral-level data for transitivity has been established. Transitivity is the basic economic idea (it is one of the axioms of expected utility theory) that a subject who prefers A to B and B to C must prefer A to C if the economic value of each item (A, B, and C) is to be preserved. It is important to economic theory because the idea of economic value depends upon it. In behavioral experiments, the monkeys’ choices do in fact exhibit transitivity. Monkeys who prefer juice A to juice B and juice B to juice C prefer A to C (Padoa-Schioppa & Assad 2008; K&G 2009). Establishing menu invariance of neuronal activity is crucial because it shows that the neuronal responses, like the behavioral responses, are stable and consistent, and therefore reflect transitivity (Padoa-Schioppa & Assad. 2008). Thus, evidence of transitivity in neuronal activity supports the idea that the values of goods are
represented in a common, comparable currency in the OFC neurons. In other words, transitivity is only possible if the neurons encode individual subjective values of goods on a single, common scale and not merely relative (menu-variant) values. Each good on offer, then, has its absolute (as it were) subjective value (SV) represented by particular neurons on a common scale.

I need to briefly review the characteristics of SVs here since their philosophical implications will be crucial to the final argument. SVs are the mean firing rates of particular populations of neurons in response to goods/items or actions. Importantly, this means that they are to be understood as cardinal numbers (i.e. they denote a quantity, in this case a neuronal firing rate, as opposed to just an ordinal position or rank). For any object that becomes an option in the context of choice, its SV is encoded in the mechanisms constitutive of valuation in that particular organism’s brain. This means that SVs are neural representations of both absolute and relative features of external items. Let me briefly detail this important philosophical point.

SVs are learned through experience. This learning process involves the phasic activation of dopamine neurons rooted in the midbrain, in particular, neurons originating from the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) and projecting to the basal ganglia and frontal cortex. In some of the initial work on valuation processes with monkeys, Schultz et al. (1993) showed that individual dopamine neurons in these midbrain dopamine pathways responded strongly to unconditioned rewards (i.e. rewards that produce a response even in monkeys not previously trained to respond to the stimulus) but not to conditioned rewards. This was some of the first evidence that these dopamine neurons encode something other than merely pleasurable experiences. Later work by Montague et al. showed that what those neurons actually encode is the difference between the monkeys’ expectations
about the reward and its actual experience of an obtained reward, as encoded in dopamine neurons. This is difference is called reward-prediction error, which we defined as:

\[
\text{RPE} = \text{df} \times (\text{the difference between the experienced outcome of an action and the forecasted outcome of that action, scaled by the rate of learning from experience, which determines the weight given to recent versus remote experiences (K&G 2009).})
\]

It is important because it helps to explain the portion of subjective values based on individual, subjective preferences.

To say that the subjective values of items or actions are partly determined by individual preferences is to say that those values are partly determined by the individual’s experience of reward, or perhaps more simply, that individual’s response to that reward. Neuroeconomists call that experience \(\text{ExperiencedR}\). \(\text{ExperiencedR}\) is the experience of a positive outcome from a reward. Rewards are objects or events that generate a particular class of behaviors (such as consummatory and approach behavior) that support basic and essential processes. Thus the experience of reward not only produces such behaviors but also positively engages with emotion systems and influences learning processes (Glimcher 2009).

The other half of subjective values, the part that represents objective features of objects or actions in the world, is expected reward or \(\text{ExpectedR}\). \(\text{ExpectedR}\) is essentially a prediction about the magnitude and probability of a reward given the previously accumulated evidence. That evidence takes the form of stored information about the physical, motivational, and economic properties of an item or action. The information is in an important sense objective: it concerns the actual physical properties of items or actions and the actual physical effects of those properties on the organism. This is one sense in which subjective values are objective.

As I argued in Chapter 5, both the objectivity and subjectivity of subjective values are inextricably intertwined in both expected and experienced reward. Somewhat crudely, we might
think of the *objectivity* of subjective values arising from the impact of physical and chemical characteristics of objects that impact upon our sensory organs and the *subjectivity* of subjective values as arising from the experience produced by the processing of that electrochemical information in our nervous systems. Because both the interplay between our sensory organs and the physical world *and* our experience of the changes effected in the nervous system figure into both expected and experienced reward, it turns out that objectivity and subjectivity figure into subjective values twice, once in the construction of *ExperiencedR* and once in the construction of *ExpectedR*. Neuroeconomists describe the firing rates of the dopamine neurons in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) of the midbrain that encode RPE as:

$$DA(\text{spikes/s}) = \alpha(\text{ExperiencedR} - \text{ExpectedR})$$

where $\alpha$ represents the scaling parameter mentioned previously that controls the rate of learning.

The crucial point for disputes about MM and the Humeanism dispute in particular is that philosophically speaking both objectivity and subjectivity figure into *both* parts of the formal definition of SV in the neural model.

Those SVs are learned through experience. The learning takes place in networks distributed in the basal ganglia and frontal cortex. More specifically, the neuronal networks implicated in the learning of values based on RPEs have their origins in the striatum, the ventral striatum in particular, and project to the frontal cortex, and the medial prefrontal cortex in particular. Much is already known about the lower-level mechanisms that constitute learning in these areas. The learning involved in valuation is made possible by synaptic plasticity, or changes in the strength of the connections between neurons. The networks of dopamine neurons disseminate information concerning learned SVs throughout these regions of the brain.
To sum up, SVs are encoded in the medial prefrontal cortex and ventral striatum of the mammalian brain. The data suggests that these areas serve as the point at which SVs are implemented into the mechanisms that produce choice, the second stage in the two-stage model.

Thus the central feature of valuation is the theoretical construct of SV and the mechanisms through which SVs are learned. In general terms, the trouble for the location project stems from a tension between the cognitive-level FP story about an agent’s subjectively valuing an item or action and the neurophysiological mechanisms upon which that story must supervene. In more specific terms, the problem for the FP realist is that (1) SVs “exist” – they are genuine, measurable neural entities, and (2) their contribution to decision and motivation processes—i.e., their explanatorily relevant characteristics and functions—pertain uniquely to the biophysical level. Here are those two points summarized by the leading neuroeconomist, Paul Glimcher:

1. “There is nearly universal agreement among neurobiologists that a group of neural systems for valuation has been identified” (2009: 511-12); and

2. There is a large body of data which supports the hypothesis that “learning mechanisms distributed through the basal ganglia and frontal cortex contribute to the construction of what we refer to as subjective value. These areas are hypothesized to learn subjective values, at a biophysical level, through the well-studied process of synaptic plasticity” (519).

I discussed that body of data in detail in Chapter 5. I now turn to two distinct challenges that it raises for the FP location project.

§6. Two Challenges for Folk Psychological Realism

A successful account of the location of FP states in the emerging scientific model of decision will have to preserve the causal relevance of beliefs and desires. This is true both as a general matter and more specifically for the Humeans, anti-Humeans, internalists, and externalists (and so on) whose theories of moral motivation depend upon the causal efficacy of
beliefs and desires. As a general matter, if we concede that neither beliefs nor desires actually *do* anything then we have removed all motivation to characterize them functionally, i.e. in terms of what they do, or for that matter in trying to reduce them to, identify them with, or otherwise suprervene them upon conglomerations of neural activity they someway or another play precisely the same causal role. This is still more obvious in the particular case of philosophical theories of moral motivation. If beliefs and desires are causally inert then all such theories are rendered empirically irrelevant for the simple reason that each identifies one or the other state or both states jointly as causally sufficient for moral motivation.

*The Challenge of Causal Relevance*

The first challenge for the location project then is to get beliefs and desires, whether construed as functional roles (e.g. Canberra Plan), sentences in the head (e.g. Fodor), metaphorical ‘directions of fit’ (e.g. Anscombe 1957 and her many followers), etc., to supervene on the immediate causes of choice selection, namely the specific cellular and molecular configurations that result from synaptic plasticity in dopamine pathways and which make possible measurements of SV, without jeopardizing their *causal relevance* in the second stage—choice—which is directly related to motor implementation.

There are several options in any choice context. What the data from neuroeconomics shows is that making a choice is a matter of selecting from among a set of actions or items—*each* of which has a SV quantifiable in neural terms—a highly valued item and implementing information concerning that option in motor pathways. For the FP realist this process of selection must involve a set of beliefs and desires about the options on offer. The trouble is that if beliefs and desires are instantiated in this account at all then they must be instantiated in such a
way as to represent the features of each of the items or actions on offer and without jeopardizing the role of beliefs and desires in the kinds of explanations for moral motivation that Humeans, internalists, and their opposition are offering.

Prima facie, the most plausible way (and to my mind the only immediately obvious way) for the FP realist to locate FP states in the neural account is to insist that desires are identical to or somehow constituted by SVs, or, again, a bit more precisely the cellular/molecular configurations that make their measurement possible. On this account of location, for any given context of choice with more than one option the FP realist will have to claim that choice involves selecting from among competing levels of desire. A monkey faced with the choice of grapes, bananas and raisins is essentially faced with the task of selecting from among competing desires for each of the fruits, and perhaps chooses on the basis of beliefs about the quantities available. Two grapes, the monkey believes, satisfy its desires better than one raisin. Dopamine, synaptic plasticity, learning and so on are merely the neurophysiological mechanisms upon which the cognitive events supervene.

Crucially, though, this approach to location appears to jeopardize the explanatory relevance of beliefs and desires in the traditional disputes about moral motivation. For example, Humeans claim that moral beliefs are insufficient for motivation because they require the presence of a desire (or similar conative state) to motivate. Anti-Humeans deny this, generally because they are drawn to some kind of motivational internalism. On the account of location just given, the

---

Some might even take the possibility of the monkey example to be an objection to FP realism in itself. The mere fact that we can conduct decision research of this kind on non-human mammals who appear to lack equivalent language capacities to those of human beings suggests that there is something seriously wrong with taking beliefs and desires so literally. Of course, one man’s Modus Ponens is another man’s Modus Tollens. The FP realist might reply that what this really shows is the absurdity of animal models of human cognition. I think that reply is severely weakened not only by the convergence with human research presented in this chapter but by the success of animal models in general.
Humean theory (or better, the spirit of that theory) will be true only trivially and its opposition simply a nonstarter. It is true in a manner of speaking that desires are required for motivation, but the point is trivial because desires are present to varying degrees in each of the options, including those that are ultimately bypassed. Moral beliefs about the nature of the possible options are insufficient for motivation because all such beliefs in the context of choice are insufficient for motivation. It is a platitude that desire (so understood) must be present for motivation precisely because in any real choice it is always present in its making a contribution to SV.

Neither does this result serve as evidence for internalist forms of anti-Humeanism that postulate besires, i.e. states which are simultaneously belief-like and desire-like, since in any given decision each of the options not chosen will be motivationally inert despite each being the object of our besires (as it were).

The rapid proliferation of FP states here—which results from our having a rather hazy commonsense conception of precisely what kinds of entities they are and consequently no principled or reliable method for picking out their realizers—prevents them from contributing anything of value to explanations of moral motivation couched in causal language. For when we

---

65 The point seems to be something of a neuroscience analog for some recent objections to the possibility of formulating a so-called “belief-desire law” which some functionalists suppose capable of explaining the relationship between intentional states in the theory-theory. Such a law might claim, for example, that “people do what they believe will satisfy their desires.” In a notable objection, Gauker (2005) argues that there are no such laws. First he criticizes the “simple formulation” of the belief-desire law according to which people do what they believe will satisfy their desires by pointing out that “there is never just one thing people desire; they always desire a lot of things. They cannot do everything they think will satisfy all of their desires, because they cannot do all of those things at once” (126). This data suggests that it is not just a platitude that people desire lots of things and cannot do them all, but that it is a fact about our neural architecture that even if we could locate some scientific analog for desire, say as part of SV, its ubiquity in the context of choice would render it explanatorily inert anyway. (I suspect, however, that this is not an argument that Gauker himself would endorse).
gloss these complex neurophysiological processes in commonsense FP terms we end up abstracting too far away from the mechanisms most immediately relevant to the explanation. Given the mechanisms of valuation, the claim that an agent chooses a particular (moral) course of action because she desires to is really just vacuous.

Thus causal relevance poses a daunting challenge for FP realism. Because the data suggests that the neural mechanism for choice selection centers on SV, the FP realist is compelled to identify desire with that particular feature of the neural story. Yet it seems impossible to straightforwardly identify desire with SV without proliferating the former to the point of rendering it explanatorily irrelevant. The realist will need some alternative account.

Two possibilities seem clear. On the one hand, she might look to develop a more nuanced account of location. Rather than straightforwardly identifying desire with subjective value, she might look to identify one particular component of SV that might serve as a supervenience base. The problem here is that the more nuanced the account becomes, the fewer commonsense cases of desiring it is likely to cover. Accurate folk psychological ascriptions of desire are not generally thought to require an intricate knowledge of neural states. On the other hand, it might seem that I have stacked the deck with my initial proposal. Why must the FP realist look to identify desire rather than belief (or belief and desire) with SV? The answer, I think, has to do with the distinguishing features of those as the realist construes them.

The Challenge of Belief-Desire Directionality

I began this argument against the location project with the rather obvious suggestion that the realist look to identify desire with subjective value. The suggestion comes naturally because it is essential to FP realism not just that we find supervenience bases for belief and desire but that
those bases, whatever they turn out to be, preserve the characteristics of those states revealed by *commonsense* description. In particular, what the FP realist must preserve is what we might call *belief-desire directionality*. This is the peculiar property of beliefs and desires that allows them to be identified in terms of their “directions of fit.” The idea seems to have originated in Anscombe (1957). In any case it is now widely used by philosophers of mind to capture the difference between beliefs and desires (and also to explain how the folk distinguish between them). Beliefs are states that depict the world as being a certain way and so are said to have a mind-to-world direction of fit while desires are states with a world-to-mind direction of fit.

Of course, some philosophers have worried about resting theories of mind on a simplistic metaphor and so have sought to cash it out, such as by using counterfactual conditionals (see Smith 1987: 54; Smith 1994), or by contextualizing it such as in debates about practical reason (see Little 1997). But whether they take the directionality metaphor literally or prefer to spell it out in more concrete terms the crucial point is that they must ultimately preserve it as part of our standard, commonsensical or folk picture of human psychology, which, as Smith puts it, “is important because it provides us with a model for explaining human action” (1997: 9). It is thus a desideratum on the account of location that it preserves belief-desire directionality because without it the FP realist loses her grip on the characteristic commonsensical nature of FP states.

Just before I suggested that on the most obvious account of location, desires are identical to or somehow constituted by SVs. The suggestion was hardly arbitrary. It is in fact directionality that makes it an obvious first (and quite possibly only) choice. It is the subjectivity of subjective value that suggests it as an obvious candidate for realizing desire, which is, the realist tells us, a subjective state with a world-to-mind direction of fit.

But now consider what role remains for belief on this account. The realist can perhaps say
that the monkey believes that each of its options carries a specific value in terms of its desirability. But this seems to conflate belief and desire in the traditional philosophical sense of the terms. Beliefs, the realist tells us, are about objective states of affairs or facts, not representations of facts about our subjective experiences of desire. This is important because the interesting question is no longer whether desire must be present in addition to belief in order for choice and motivation to occur—which is the question contested by Humeans and anti-Humeans—but rather just the opposite: how beliefs about the desirability of an action or item contribute to choice selection.

To see this, consider first that it is a consequence of the moral motivation framework and Humeanism in particular that we must find some role for both belief and desire in the neural explanation. Finding neural correlates for desire at the cost of preserving any role for belief in a neural explanation for moral motivation is hardly a victory for FP realism. But while it seems clear enough that for the FP realist desires must somehow be closely connected to SV, it is far less clear what role remains for belief, except perhaps to say that agents have beliefs about their subjective desires (SVs). This, though, yields the peculiar result that an agent navigates the world using representational desires and that her course of action is ultimately determined by the presence of a scale-tipping belief about which is the optimal desire to satisfy. That is, this particular account of location might find some room for both belief and desire only by turning the dispute about Humeanism in the wrong direction. Thus the challenge of causal relevance compels the realist to locate desire inside the neural account of subjective value, while such an approach tends to jeopardize belief-desire directionality.

Finally, some, such as Schroeder et al. (2010), might conclude that these considerations about belief-desire directionality support or are at least compatible with instrumentalism about
moral motivation. This again is the view that an agent is motivated when she forms a belief about how to satisfy her pre-existing desires. Initially this seems a tempting conclusion, particularly given that the fundamental difference between instrumentalism and standard-form Humeanism is directionality: the Humean claims moral beliefs require the presence of a desire to motivate while the instrumentalist claims that motivation involves forming beliefs about how to satisfy prior desires. I remain doubtful for at least two reasons.

First note that instrumentalism will continue to face the challenge of causal relevance in virtue of its commitment to the causal efficacy of FP states. At best then instrumentalism will be poised to handle only one half of the problems that it inherits from its FP ancestry.

Second, and more importantly, I doubt that it solves even that much. For what this discussion of directionality shows is not merely that philosophical theories of moral motivation have got things backwards but that at best they have got things backwards. Before we can even worry about how to preserve directions of fit we must first establish that we can plausibly locate beliefs and desires in the neural account without endangering folk expertise on believing and desiring. But preserving traditional FP states by construing beliefs as the representations of our desires does just that. Beliefs on the commonsense picture are supposed to be about the world, not representations of our subjective desires.

The possibility remains that my initial proposal—somehow relating subjective value to desire—was simply too general. Perhaps it would be more charitable to look for a way to locate beliefs and desires in the two-stage model more directly. We can take on this question by returning to a more technical formulation of valuation and subjective value.
§7. Beliefs, Desires, and Neural Architecture

Following Paul Glimcher (2009; 2010) we can understand subjective values (SVs) as theoretical entities with the following features. They are neural representations of the reward value of an action or object as determined by our subjective preferences and the objective features of a reward (e.g., its quantity). They are defined as the mean firing rate of a specific population of neurons, i.e., the “common currency” of neuroeconomics. They are thus real numbers ranging from 0 to 1000 and their units are action potentials per second. These values are purportedly learned through iterative updating with experience. Extensive research indicates that this is achieved through dopamine (DA) reinforcement learning. More formally, the subjective value of a single option $j$ (in a context of choice) is represented as:

$$SV_j = \sum_i w_i x_{ij} / \sum_i w_i$$

where $i$ indexes each of the brain’s neurons, $x_i$ is the firing rate of the $i$th neuron, and $w_i$ is a weight (ranging from 0 to 1) which describes the additive contribution of that particular neuron to the SV of $j$. In other words, the subjective value of some particular object or action $j$ (e.g. a food or a turn of a handle in an experiment) is represented as the average weighted firing rate of a subpopulation of neurons which does the encoding (2009: 511). For example, suppose $j$ is an action, say the turning of a handle in an experiment with a monkey. In this case the region of the brain that encodes the action, such as the superior colliculus (SC)—which encodes action in a topographic map as previously described—will contain cellular activity in a restricted area of the map that encodes the value for $j$.

In practice the regions of the mammalian brain that contain the neuronal circuitry capable of giving us the mean neuronal firing rates that might plausibly encode subjective value have already been identified. As we have seen, these are the striatum and the medial prefrontal cortex.
in virtue of their being regions which satisfy two theoretical desiderata: (1) data indicating a firing rate pattern, such as BOLD activation in fMRI, linearly correlated with the utility of actions or objects and (2) data indicating a compact population of neurons capable of maintaining the linear correlation with the subjective value of \( j \) (Glimcher 2009: 511; see section 4.4 above for the review of that data).

In the above formula \( w \) is a weight ranging from 0 to 1 describing the additive contribution of a particular neuron to the SV of an object or action. The empirical question upon which the successful localization (i.e. physical realization) of a valuation system hinges is whether there is a population of neurons that can provide the non-zero values for \( w \). One of the keys to answering this question is to determine whether there are discrete populations of neurons in brain areas whose activation correlate with SV in reward and choice conditions (Glimcher 2009: 514). Surprisingly, the data in the previous sections from work on the ventral striatum and medial PFC seem to indicate that the answer is yes.

It seems, though, that this last empirical question takes us one step farther than we need. For the question with which we began was not whether we should expect to find physical realizers for the folk roles of belief and desire but only whether we should expect to find “theoretical realizers” whose physical counterparts might eventually be determined by some future neuroscience. Where in this technical formulation do we, or could we, locate empirical theory counterparts for the relevant folk roles?

In the last section I said that the most plausible answer was somewhere in the theoretical construct of subjective value. The problem the FP realist faces here is that even if these neuroscientists have wrongly identified the striatum and medial PFC as the physical cell populations which implement the valuation mechanism in the human brain, i.e., even if we are
wrong about the physical realizers, the above formula provides no obvious candidates for theoretical realizers. The summation of neural firing rates of whatever population of cells and the weighted contribution of some particular neuron to the firing rate of some population of neurons are bad candidates for supervenience bases precisely because none of the commonsense observations about the properties of belief and desire in everyday discourse—the observations from which we generate the platitudes on the functionalist account—are consistent with the properties of the entities represented in the above formula for SV.

Recall that in Chapter 3 I argued that the mistaken assumption embedded in Canberra-style moral functionalism is that because we know that some function must occur to bring about a phenomenon (and that some physical thing must make that function possible) that we know anything about how the phenomenon is brought about, i.e., that we must know anything about the nature of the function that brings about the phenomenon. Here the idea is more concrete. The more specific nature of the functions of beliefs and desires qua functional roles seem to be rather different from the nature of the physiological functions that bring about subjective value representations.

Another source of support for this claim, I think, turns out to be an updated version of Stich’s complaint with the theory-theory, which I discussed in Chapter 3. According to Stich, one of the fundamental tenets of folk psychology is that a belief is a state that can interact in many ways with many other states. Consequently, it can be implicated in the causal history, or etiology, of a variety of behaviors (230). According to Stich the problem is that the empirical data suggests that this fundamental assumption of FP is a bad one.

Here the neuroscience analog to Stich’s argument from cognitive science emerges: to save the distinction between beliefs and desires in the face of this empirical data on decision we need
to find a physical or theoretical basis for the distinction. If beliefs and desires are explanatorily relevant to moral decision-making, they had better not turn out to be indistinguishable features of the same process because this runs counter to what we get from commonsense platitudes about folk practice.

Suppose we want to associate one set of neurons with one intentional state and another set of neurons with the other. How in the theoretical terms presented above do we capture the distinction between cells that represent the world as it is and cells with which the world must be made to fit, so to speak? In fact I think that with all of this in the background we can attach some neuroscience to the end of Stich’s argument (which, incidentally, was presented more than two decades ago) concerning intentional psychology and cognitive science. Here is the relevant passage quoted in full:

For folk psychology, a belief is a state which can interact in many ways with many other states and which can be implicated in the etiology of many different sorts of behavior... But, of course, the belief that p does much more than merely contribute to the causation of its own linguistic expression. It is a fundamental tenet of folk psychology that the very same state which underlies the sincere assertion of ‘p’ also may lead to a variety of nonverbal behaviors. There is, however, nothing necessary or a priori about the claim that the states underlying assertions also underlie nonverbal behavior. There are other ways to organize a cognitive system. There might, for example, be a cognitive system which, so to speak, keeps two sets of books, or two subsystems of vaguely belief-like states. One of these subsystems would interact with those parts of the system responsible for verbal reporting, while the other interacted with those parts of the system responsible for nonverbal behavior. Of course it might be the case that the two belief-like subsystems frequently agreed with each other. But it might also be the case that from time to time they did not agree on some point. When this situation arose, there would be a disparity between what the subject said and what he did. What is striking about the results I shall sketch is that they strongly suggest that our cognitive systems keeps two sets of books in this way. And this is a finding for which folk psychology is radically unprepared (231).

Stich follows this with a review of empirical evidence from social psychology intended to undermine two assumptions made by ordinary belief ascription: (1) an assumption about the general organization of our cognitive architecture, namely, that we have belief-like states which
interact with desire-like states as in a strong representational theory of mind (STM) and (2) an assumption of modularity in the organization of our belief, i.e., that there is some isolatable part of the system that plays the central role in a typical causal history leading to the utterance of a sentence. But we can now add to Stich’s list the new data from neuroeconomics. The organization of not just our cognitive architecture but our neural architecture when it comes to weighing options and making decisions seems to indicate that there is just no good way to preserve the needed theoretical distinction. And this too is a finding for which folk psychology is radically underprepared.

I set up my argument by claiming that my goals were rather more modest than Stich’s. Rather than try to defend eliminativism, I just wanted to show that wherever beliefs and desires might appear in the neuroscience, they do not appear in the relevant ways or in the relevant capacities to preserve the value of the Humeanism dispute. My point in taking this line is to concede one of the more controversial parts of Stich’s project. I see no special benefit in denying that we can find some way to make sense of belief and desire in neuroscientific research precisely because neuroscientists are just members of the folk who, when working at the relevant levels of description, trade in these terms the same as everyone else. In particular, I think this is true at the behavioral level. Even though there seems little hope of preserving the explanatory relevance of beliefs and desires at the level of valuation and choice, it is quite useful to say that a monkey believes he has a choice between grapes and bananas and that perhaps he desires one more than he desires the other. Further, it is admittedly hard to see how the two-stage model ever gets off the ground without this foundational behavioral research (cf. Churchland 1986). And in that very basic sense, there is no special reason to push the claim that beliefs and desires appear nowhere in neuroscience.
The mistake is to go along with the popular idea that because a case can be made for the presence of intentional psychology *somewhere* in the neuroscience community that it must therefore be relevant to the future neuroscience of decision. The most interesting thing to emerge from all of this empirical literature is the very basic observation that scientists work at different levels, some of which overlap, and some of which are more directly relevant to the mechanistic explanation of a phenomenon than others. Researchers who use fMRI scans might be convinced that consistent activation of some brain region in a cognitive task indicates that somewhere within that region a mechanism is occurring that brings about the cognitive phenomenon in question. But we cannot on that basis alone insist that the causal mechanism itself has been identified. The data is correlational. In some minimal sense we might think that the relevant mechanism makes an appearance, in the form of some poorly understood contribution to rCBF or the BOLD signal, but it would be a mistake to claim that the data therefore elucidates the mechanism.

Finally, one might object that the neuroscience as presented here just has the theory all wrong. Perhaps future neuroscience will prove unkind to neuroeconomics and its two-stage decision model, for example. Folk psychology, unlike so many newfangled scientific research programs, has withstood the test of time. Though the observation is reasonable, the problem with hanging an objection on it is that it depends on the unlikely outcome that the results that eventually emerge from some future neuroscience will be different in *kind* rather than merely in degree from those presented here. If the extensive convergence of data from the last fifty or so years of work in neuroscience considered here has shown anything, it has at least shown that scientific progress isn’t entirely haphazard. Science often builds up from a set of initial hypotheses, which get revised and reformulated along the way.
No doubt many of the claims emerging from decision science and neuroeconomics will need to be revised, some quite substantially, and some even discarded. It is even plausible to imagine that the project might eventually be abandoned altogether. Even so, what we lack is any good reason for thinking that the research programs that eventually succeed it will move away from studying the neural mechanisms—the activity of specific neural populations and the contributions of individual cells to regional activation patterns—that serve as proximate explanations for behavior. When an animal or person acts in accordance with a judgment or decision, it does so because cells in the central nervous system conduct electrical impulses to the animal’s muscles via axons that project outward to the autonomic nervous system. Working backward from motor activity, then, is a defensible strategy. A good, proximal explanation for why Moral Agent S did what she did will have to tell a story that ends with cellular signaling of just this sort. And the question we are ultimately after in disputes about motivation and behavior, including moral motivation, is how it is that Agent S came to act as she did. There may very well be cognitive-level explanations (or indeed some other explanation entirely) for what in turn brings about these cellular signaling processes. But that such further explanations may be forthcoming is no objection to working backward from the most proximate cause of the skeletomuscular system activity that we call behavior. That is what I have tried to do here. Those who wish to persist in the idea that the data presented here is irrelevant owe a more substantive account of its shortcomings. But even that is a project that we should not expect to pursue adequately from the armchair.

The time is ripe to begin rethinking the commonsense psychological framework upon which contemporarily analytic ethics is built. This goes especially for self-described naturalists. Genuinely empirical approaches to moral philosophy must now begin to address the dubious
psychological assumptions implicit in their traditional philosophical theories rather than merely use scientific data to adjudicate between them.
References


