ABSTRACT

Title of Document: THE DISUNITY OF MORAL JUDGMENT: AN ESSAY IN MORAL PSYCHOLOGY

Leland Frederick Saunders, Ph.D. 2011

Directed By: Professor Susan Dwyer, Philosophy

Recently there has been great deal of interest in uncovering the psychological processes of moral judgment. Over the past 10 years, psychologists and neuroscientists have studied the psychology and neurology of moral judgment, and there are now several empirical models of the psychology of moral judgment that attempt to explain these empirical findings. I argue that current empirical models of moral judgment, however, are inadequate, because the psychological processes that they posit cannot explain some important characteristics of other features of our moral psychology. On the other hand, contemporary philosophical accounts of moral judgment do not fare any better, because they are not consistent with recent empirical findings.

My diagnosis for these inadequacies is that contemporary philosophical and empirical models of moral judgment are implicitly committed to what I call the Unity of Process Thesis, which is the claim that all moral judgments are the products of a single psychological process. I argue that the Unity of Process Thesis must be abandoned, because it makes it impossible to account for some important features of our moral

psychology. What is needed is a dual-process model of moral judgment, and by drawing on an empirically well-supported dual-process architecture of human judgment, I develop a framework for moral judgment that posits two distinct kinds of moral judgments, intuitive and deliberative, that have very different underlying psychologies that operate in different ways, using different cognitive resources, that are tied to motivation in different ways, and play different roles in our moral psychology. I call this framework the Two Kinds Hypothesis.

The distinction between intuitive and deliberative moral judging and judgments is quite valuable in developing an overall psychological picture of moral judgment that captures important features of our moral psychology and that is consistent with current accounts of the general architecture of human judgment. This analysis also has upshots in illuminating some debates in metaethics as well, specifically the debate between moral particularists and generalists, and the debate between moral judgment internalists and externalists.

THE DISUNITY OF MORAL JUDGMENT: AN ESSAY IN MORAL PSYCHOLOGY

By

Leland Frederick Saunders

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctorate of Philosophy
2011

Advisory Committee:
Professor Susan Dwyer, Chair
Professor Peter Carruthers
Professor Dan Moller
Professor Rachel Singpurwalla
Professor Peter Mallios

Dedication

For Correne

Acknowledgements

I would like to thank the tireless work of my committee in helping me bring this project to completion. In particular, I want to thank Susan Dwyer for her many, many comments on many, many drafts. Her probing questions always pushed me to develop better arguments and to express my thoughts more clearly. Moreover, she was the one who first introduced me to the central topics in this dissertation in a graduate seminar I took with her in 2004. I have her to thank for my interest in moral cognition and moral psychology. I would also like to thank Peter Carruthers, who first introduced me to dual-process theories of cognitive architecture. Many of the ideas in this dissertation grew out of a paper I wrote for his seminar on the architecture of the mind. That paper was eventually published, and has now become a dissertation. His skeptical but charitable approach to my work helped to make it much better. He also provided many, many comments on many, many drafts. Aside from these two individuals, there are countless others who have helped me think through the issues in this dissertation in more informal settings. I would like to thank, in particular, Matt King, Ryan Fanselow, and Ryan Millsap for our many conversations on these topics.

However, none of this would have been possible without the support of my wife, Correne, to whom this dissertation is dedicated. Our life is far more complicated now than when I began, thanks to our two sons, but even so, she has always provided steady encouragement and support. I cannot thank her enough.

Table of Contents

List of Tables

List of Figures

Chapter 1: Moral Psychology & Moral Judgment

Moral psychology involves the investigation of a broad and complicated set of issues at the intersection of human psychology and moral theory, including, among others, whether and how personal projects, plans, and commitments create areas of moral concern; the structure of moral motivation, moral reasoning, and moral decision-making; the psychological processes of moral judgment; and the features of moral agency. While the central areas of concern in moral psychology involve the intersection of psychology and moral theory, it is important to distinguish between two important, but distinct projects in moral psychology: the psychological project, and the normative or epistemic project (Held, 1996). The psychological project in moral psychology aims to uncover the psychological mechanisms and processes that explain how it is humans engage in moral practice; i.e., it attempts to explain, among other things, how humans reason in the moral domain, how moral motivation is psychologically realized, and the psychological processes of moral judgment. The psychological project in moral psychology is thus primarily a descriptive one in that it aims to describe the actual psychological processes that subserve human moral practices. The normative project in moral psychology, on the other hand, is centrally concerned with recommending certain decision procedures, moral theories, moral attitudes, and moral judgments; i.e., it attempts to specify the normative requirements that ought to govern one's moral thinking and acting.

Though there are two distinct projects in moral psychology, arguments and findings in one project are often used to resolve, dissolve, or illuminate disputes in the

other. For example, psychological arguments have been used to undermine the plausibility of particular moral theories (Doris, 2005; Frankfurt, 1988; Harman, 1999; Williams, 1973a), to support a particular conception of moral agency (Korsgaard, 1996), or to advance theories of rational action (Gibbard, 1990; Kant, 1785/1996; Williams, 1981). On the other hand, normative theories have also been used to undermine the plausibility of certain psychological claims (Fine, 2006; Kennett, 2006; Kennett & Fine, 2009), or to advance a particular view of the psychology of moral judgment (Herman, 1993; Smith, 1994). There are deep and interesting questions with respect to how these two projects in moral psychology bear on one another (see, for example, Flanagan, 1991; Held, 1996), but even so it is important to keep a clear grip on this distinction.

This dissertation is primarily concerned with a question in the psychological project of moral psychology that has received a great deal of attention recently, namely, what are the psychological mechanisms and processes that lead to a moral judgment? The psychology of moral judgment has been the focus of increased attention recently. In part, this is because the psychology of moral judgment figures centrally in a number of important metaethical debates, such as whether moral judgments are cognitive or non-cognitive mental states; and whether moral judgments necessarily motivate. But another reason for this recent interest is that uncovering the psychological processes of moral judgment seems to be directly amenable to ordinary methods of empirical investigation, and over the past 10 years psychologists and neuroscientists have amassed significant body of empirical literature with respect to the psychology and neurological correlates of moral judgment; all of which promises

to shed new light on old problems by providing insights into the psychological workings of moral judgment that cannot be discovered through purely philosophical methods. This is an exciting time to be working in moral psychology, and in working out the psychological processes of moral judgment, but so far enthusiasm has outstripped precision. Empirically informed models of moral judgment abound, but there has been little sustained attention to the question of what models of moral judgment are meant to explain and what, if any, explanatory and theoretical constraints are central in assessing them.

This is a conspicuous oversight, and one of central aim of this dissertation is to provide a sustained investigation into these questions. This is important, not only because getting a clear understanding of the psychological processes of moral judgment is interesting in its own right, but also because many contemporary moral psychologists attempt to draw significant and skeptical normative conclusions from their respective models of moral judgment. Many contemporary moral psychologists argue that the psychological processes of moral judgment show that moral judgments are not, and cannot be, based in reasons, and thus that moral judgments are not rationally assessable in any meaningful way and that morality itself is not a rational enterprise (Haidt & Joseph, 2007; Haidt, 2001; Haidt & Bjorklund, 2008a; Haidt & Joseph, 2004; Joyce, 2006; Prinz, 2007). Determining whether such skeptical claims are warranted requires, in part, determining whether there are good empirical and philosophical reasons supporting those models of moral judgment that are supposed to give rise to them. And doing that first requires getting clear on what a model of

moral judgment is supposed to be, and what constraints are central in assessing them. That is why such a sustained investigation in warranted.

The central aim of this chapter is to lay the groundwork for this investigation by outlining what it is a model of moral judgment is supposed to be, and what general constraints should guide theorizing with respect to the psychology of moral judgment. A quick terminological point is in order before moving on: the term "judgment" is often used to refer to three distinct notions: judgment as an activity or psychological process; judgment as a product or result of that activity or process; and judgment as the content of a mental state produced by that activity or process. When referring to judgment as an activity or psychological process, I shall use the locution, *moral judging*. When referring to judgment as the output of these processes, I shall use the term, *moral judgment*. To refer to the content of a moral judgment, I shall talk of what the person *judged*, as in whether a person judged some action to be permissible or impermissible, etc.

1. What Is a Model of Moral Judging?

A model of moral judging is supposed to provide an account of the psychological mechanisms and processes that lead to a moral judgment such that it explains the capacity of every ordinary human being to judge the morality of actions, events, policies, or persons. All ordinary humans morally evaluate the actions, events and persons around them, sometimes without any conscious awareness of doing so. Making such judgments requires that minds like ours move, sometimes imperceptibly (though sometimes with considerable thought), from descriptive representations (real or hypothetical) of actions and persons to normative conclusions (moral judgments)

with respect to them. It is our capacity to do this that gives rise to a central question in moral psychology: what is it that explains how minds like ours are *capable* of producing moral judgments on the basis of non-moral representations? It is this question that a model of moral judging is supposed to answer. It is supposed to provide an explanation of this capacity.

Of course, not all explanations are the same, and different kinds of explanations can be aimed at different levels of abstraction. Marr, for example, distinguishes between three levels of explanation for a cognitive system: computational, algorithmic, and implementational (Marr, 1982/2010, pp. 22-27). An explanation aimed at the level of computation is the most abstract, and it specifies what the goal of a computational system is (such as a capacity), what information gets computed, and how information flows within the system. An explanation aimed at the level of an algorithm specifies how the how computations are implemented in an algorithm, which specifies the precise computations that transform information within the system from input to output. An algorithmic explanation is less abstract than a computational explanation, but more abstract then an explanation of implementation. An explanation aimed at the level of implementation specifies how the algorithm is physically realized in the operations of the brain. This is the least abstract explanation of a capacity.

Contemporary models of moral judging are aimed at the computational level of explanation: they are supposed to explain the informational processes that lead to a moral judgment. This is usually done by providing a "boxology"—a diagram that represents the flow of information within a system using boxes and arrows. The boxes

in a "boxology" represent discrete informational processes, and the arrows between the boxes represent causal relationships. Such "boxologies" are similar to what Cummins calls a functional analysis (Cummins, 1975, 2000). A functional analysis breaks the capacity of interest, in this case, the capacity for moral judgment, into constituent capacities that are both less sophisticated than, and different in kind from, the capacity being explained. This provides a computational explanation for how the capacity as a whole operates: information is computed as a programmed exercise of the underlying analyzing capacities. As Cummins writes: "Functional analysis consists in analyzing a disposition into a number of less problematic dispositions such that programmed manifestations of these analyzing dispositions amounts to a manifestation of the analyzed disposition…[where programed means] organized in a way that could be specified in a program or flowchart" (pg. 125).

It is important to point out that a functional analysis of a capacity is more than a *mere* redescription of that capacity in other terms. A functional analysis does indeed redescribe the target capacity, but it is an informative redescription—it provides an account of what psychological processes are involved, how they are ordered, and their causal relationships. Moreover, by specifying particular discrete informational processes and causal relationships, functional analyses make predictions that can be empirically tested. Thus, a functional analysis is more than a mere redescription. It is an informative and testable explanation for how information within a capacity is processed.

In short, then, a model of moral judging is supposed to be a computational explanation of the ordinary human capacity to produce moral judgments. With that in

place, the question now is what general constraints operate as conditions of adequacy on any model of moral judging. I shall focus on two in this chapter, which I take to be the two primary constraints on any model of moral judging, though there may be others.

2. The First Condition of Adequacy

The first condition of adequacy on any model of moral judging is that it must be consistent with, or better, help explain, the empirical findings with respect to moral judging and judgment. This constraint derives its force from the fact that a model of moral judging is supposed to explain the actual psychological mechanism and processes of moral judging. They are thus empirical hypothesis and can and should be tested against empirical findings with respect to moral judging. Moreover, because models of moral judging are an informative redescription of the capacity for moral judgment, they make predictions that can be tested against empirical findings. Among the most relevant empirical findings for this purpose are those of psychological effects, which provide evidence concerning what sorts of tasks, inputs, or conditions affect the output of a given capacity. These psychological effects can help decide between two competing models of moral judging if one functional analysis predicts (or is at least consistent with) the observed psychological effect and the other is not. For example, if one had to choose between two different functional analyses of moral judging, one that posited an analyzing capacity unique to the moral judging, such as a Moral Faculty (Dwyer, 1999, 2006, 2009; Hauser, 2006; Mikhail, 2007, 2009), and a second that claimed that the analyzing capacities for moral judging were a kludge of non-moral capacities (Haidt, 2001; Haidt & Bjorklund, 2008a; Prinz, 2007; Stich,

2006), discovering some set of psychological effects could help decide between them if one sort of functional analysis is consistent with the observed psychological effects and the other is not.

Thus, empirical findings with respect to moral judging play an important role with respect to constructing and evaluating models of moral judging. They can help decide between competing models of moral judging if one model is consistent with the observed psychological effects while the other is not. Thus, the first condition of adequacy is that a model of moral judging must be consistent with the empirical findings with respect to moral judging and judgment, though it would be better if a model of moral judging could help explain how the observed psychological effects arise.

3. The Second Condition of Adequacy

The second condition of adequacy on any model of moral judging is that it must be consistent with, or better, help explain, other manifest phenomena of our moral psychology, including various features of moral motivation, moral reasoning, moral decision-making, the fact that people's moral outlooks change over time, and our ordinary justificatory practices. The force of this constraint is less obvious than the first, but it is still straightforward. Any empirical hypothesis can be tested against a wide array of facts, not just those it initially seeks to explain. This is true in the case of moral judging as well. A model of moral judging is an empirical model, and as such any model of moral judging can and should be tested against a wider range of facts than those that it is initially meant to explain. What a model of moral judging is

supposed to explain is our capacity to produce moral judgments, but those proposed explanations can be tested against other facts of our broader moral psychology.

It is difficult to pin down, in advance, what features of our broader moral psychology will be relevant for testing any particular model of moral judging. It will depend upon the sorts of implications a given model of moral judging has with respect to those broader features of our moral psychology. Moreover, it is difficult, in advance, and without argument, to determine just what set of facts are that constitute the facts of our broader moral psychology. Nonetheless, if it can be shown that a model of moral judging is inconsistent with some important facts of our moral psychology, then that is sufficient grounds to challenge the plausibility of that model of moral judging.

Why should that be the case? Simply put, because the capacity for moral judgment is simply one of many constituent components of our moral psychology. As a scientific methodology, carving up the topic of moral psychology into smaller, constituent components for sustained investigation is potentially quite fruitful. However, it is important to recognize that studying moral judging as a constituent of moral psychology involves abstracting it away, in some degree, from other aspects of our moral psychology, and idealizing its functions in certain ways that ignore some parts of the complex set of factors that lead to an individual moral judgment. Even though this is methodologically useful, in the end a model of moral judging must not float *entirely* free of the complexities of actual moral practice. In the end, the theoretical and explanatory value of a particular explanation must be determined by how well it fits into the larger domain of inquiry of which it is a piece. As such, it is a

condition of adequacy on any proposed model of moral judging must be consistent with the broader facts of our moral psychology.

4. Preview

Having made some initial distinctions and developed two conditions of adequacy on any model of moral judging, I can now lay out the overall structure of the argument of the rest of this dissertation. In Chapter 2 I begin by laying out the empirical findings that psychologists and neurologists have been collecting over the past ten years with respect to moral judging. There is a wide range of empirical findings available, but two interesting findings, which are the central motivations for most contemporary models of moral judging, are that there is (1) a dissociation between people's moral judgments and the justifications people offer for them, and (2) a tight connection between moral judging and the emotions. How best to explain these findings is in some dispute, and various empirically minded moral psychologists have developed very different models of moral judgment. What they all agree on, however, is that one historically influential model of moral judging is inconsistent with these findings, namely, the deductive model of moral judgment. According to the deductive model of moral judgment, moral judgments are the conclusions of conscious deductive arguments from first principles of morality. This model of moral judging has a long history in both philosophy and psychology, and contemporary moral psychologists argue that recent empirical findings show that it is empirically defective, and thus that a new model of moral judging is needed. Only two such models are developed in enough detail to evaluate in any meaningful way: Haidt's Social Intuitionist Model, and Prinz's Constructive Sentimentalism.

Both Social Intuitionism and Constructive Sentimentalism make two claims. First, they both claim that emotions play the central causal role in moral judging. And second, they both claim that distinctively moral reasoning—that is, reasoning with moral content—does not have any direct influence on moral judging. These two claims, they argue, follow directly from recent empirical findings. This claim, however, raises two important questions: (1) does recognizing that emotions have a central causal role in moral judging *require* giving up a central causal role for distinctively moral reasoning; and (2) can such a view of moral judging, which gives up a central causal role for distinctively moral reasoning, that is, moral reasoning with distinctively moral content,[1] satisfy the constraint that a model of moral judgment must be consistent with other important features of our moral psychology?

In Chapter 3 I argue that the answer to the second question is "no," by looking at the phenomenon of moral change. Moral change is simply the observation that people's moral commitments, attitudes, and judgments can change over time, sometimes as a result of conscious reasoning. Racists, for example, come to disavow their racist attitudes, and meat eaters come to the view that eating meat is morally wrong. There is now some research on moral change, which indicates that the most straightforward explanation for some instances of moral change is that it occurs when a person reasons with distinctively moral content and comes to a considered judgment of some person, action, or practice quite distinct from his or her initial moral judgment. Both Social Intuitionists and Constructive Sentimentalists, however, must explain moral change in some other way, because neither allows that distinctively

[1] I use the locutions "distinctively moral reasoning" and "reasoning with distinct moral content" interchangeably. Both of these locutions refer to reasoning with moral concepts, such as RIGHT, WRONG, PERMISSIBLE, etc.

moral forms of reasoning can have any direct influence on a person's moral judgments. The accounts that Social Intuitionists and Constructive Sentimentalists provide come with some costs to both explanatory and theoretical adequacy. Neither can provide a fully satisfactory account of the causes of moral change. Moreover, the arguments that Social Intuitionists and Constructive Sentimentalists offer to show that distinctively moral reasoning cannot have any appreciable influence on moral judgment are not very powerful, because they both take as their primary target a form of reasoning that is simply not possible for limited epistemic creatures such as ourselves. Once one recognizes the limitations of our capacities for distinctively moral reasoning, their arguments lose much of their force. I conclude the chapter by developing a naturalized account of moral reasoning, consistent with the empirical research, and consistent with our limited epistemic capacities that can easily explain moral change.

Chapter 4 is a continuation of the themes developed in Chapter 3. Both Social Intuitionists and Constructive Sentimentalists argue that their respective models of moral judging show directly that moral judgments are not appropriate targets of rational criticism, and that moral judgments cannot be justified. They both employ a version of what I call the Regress Argument to arrive at this conclusion. However, I argue that the Regress Argument is not successful, because it involves an equivocation between to different sense of a reason. Moreover, the epistemic conclusions of Social Intuitionists and Constructive Sentimentalists are simply inconsistent with some aspects of ordinary moral experience. However, Social Intuitionists and Constructive Sentimentalists are right in claiming that there is an

important relationship between psychological claims with respect to moral judging and epistemic claims with respect to the justifiability of moral judgments; it just is not the one they think that it is.

In large part, I argue this is because Social Intuitionist and Constructive Sentimentalists have the wrong picture of the structure of moral reasoning and moral justification. They both envision moral justification as requiring deduction from first principles, but that is not the case. Reflective equilibrium, which is a back and forth between reasoning and intuition, is one account of moral justification that is quite possible for cognitively limited creatures such as ourselves, and it is consistent with the naturalized account of moral reasoning I develop in Chapter 3.

Taken together, the arguments in Chapters 3 and 4 show that the account of naturalized moral reasoning I develop accounts for certain features of our broader moral psychology much better than the views of moral reasoning offered by Social Intuitionists and Constructive Sentimentalists. This answers the second question at the end of Chapter 2, it is not possible to give up on an important causal role for moral reasoning and account for broader features of our moral psychology without important losses to theoretical and explanatory adequacy. Now the argument can turn to address the first question at the end of the chapter: does recognizing that emotions have a central causal role in moral judgment require that we give up a central causal role for distinctively moral reasoning? Again, the answer must be "no," but why do Social Intuitionists and Constructive Sentimentalists think that it does?

Chapter 5 offers my diagnosis, which is that they are both implicitly committed to the Unity of Process Thesis. The Unity of Process Thesis is the claim

that all *genuine* moral judgments are the products of a single "core" psychological process. Under the constraints of this assumption the theoretical space for explaining moral judging is seen as an opposition between reasoning and emotions. And while both Social Intuitionism and Constructive Sentimentalism are nominally dual-process models of moral judging, they are nonetheless committed to the Unity of Process because they both hold that only one psychological process is capable of producing genuine moral judgments: the emotions. But it is not just Social Intuitionists and Constructive Sentimentalists who are committed to the Unity of Process Thesis—their critics are as well—only they hold that the single "core" psychological process of moral judgment is reasoning. However, the cognitivist views of the critics of Social Intuitionism and Constructive Sentimentalism (and many moral philosophers as well) do not satisfy the first constraint on any plausible model of moral judgment; that they must be consistent with the empirical data. On the other hand, Social Intuitionism and Constructive Sentimentalism do not satisfy the second constraint on any plausible model of moral judgment; that it must be consistent with broader features of our moral psychology. A new model of moral judgment is needed that can satisfy both.

Drawing on dual process accounts of reasoning and judgment, I develop a genuine dual-process account of moral judgment that distinguishes between two types of moral judging, intuitive and deliberative, that are underpinned by different cognitive architectures and require different functional analyses. I do not fully develop functional analyses for either intuitive or deliberative moral judging—much more research is needed to do that, but the framework is still valuable for getting a

grip on a real psychological division. Moreover, given the different causal etiologies of intuitive and deliberative moral judgments, and the different way these judgments function in thinking and acting, I argue that should be considered distinct psychological kinds. Thus, I call my framework the Two Kinds Hypothesis. The distinction between intuitive and deliberative moral judgments is quite valuable in developing an overall psychological framework for studying moral judging and judgment that captures important features of moral psychology and that is consistent with current accounts of the general architecture of human judgment. According to the view I develop, there is a complex interplay between intuitive and deliberative moral judging, but these two processes operate in different ways, using different cognitive resources, are tied to motivation in different ways, and play different roles in the mental economy. Moreover, the model of moral judging that I develop is consistent with, and helps explain, the empirical data with respect to moral judging judgment, and it is consistent with, and helps explain, other features of our moral psychology, including moral change and moral justification.

As I said, one reason moral judging is such an important topic is that it figures centrally in a number of important metaethical debates, such as the debate between moral judgment internalists and externalists, and the debate between particularists and generalists. Chapter 6 shows how the Two Kinds Hypothesis can provide a way forward in these seemingly intractable debates. I do not propose to settle these debates, but instead, to provide a new way of understanding them, and perhaps, eventually for solving them. Just as many moral psychologists wrongly hold to the Unity of Process Thesis, I argue that many debates in metaethics are hindered because

metaethicists implicitly hold to the Unity of Kind Thesis with respect to moral judgments—that there is a single kind of moral judgment that always plays the same role in the mental economy. I argue that indexing universal claims with respect to moral judgments to either intuitive or deliberative moral judgments can provide a way forward in these debates.

Chapter 2: The Science of Moral Judgment

If one constraint on any model of moral judging is that it is consistent with empirical findings, then it is necessary to look at the relevant experimental findings as a starting point for theorizing, including recent findings with respect to moral dumbfounding, behavioral studies, brain imaging studies, and studies on psychopaths. Over the past 10 years, psychologists and neurologists have been investigating the underlying psychological processes and neurological correlates of moral judging, much of which indicates that there is a dissociation between people's moral judgments and the justifications they offer for them, and also that there is a tight connection between the processes of moral judging and the emotions. How to explain these findings is in some dispute, and various empirically minded moral psychologists have developed very different models of moral judging. What they all agree on, however, is that the deductive model of moral judgment, which views moral judging as a process of deduction from moral principles, is wrong. Many empirically minded moral psychologists have developed alternative models of moral judging, but only two are developed in enough detail to evaluate in any meaningful way: Haidt's Social Intuitionist Model, and Prinz's Constructive Sentimentalism. The aim of this chapter is to summarize the relevant empirical literature with respect to moral judging, and to outline the Social Intuitionist and Constructive Sentimentalist models of moral judging.

1. The Deductive Model

Some might wonder why the scientific study of moral judging is useful in the first place, because the processes of moral judging seem to be fairly introspectively obvious: moral judging is a process of deductive reasoning from first principles of morality (such as the Categorical Imperative or the principle of utility), or other, more fine-grained moral principles (such as stealing is wrong) to a moral conclusion. Call this the deductive model of moral judgment. On this view, a typical instance of moral judging involves reasoning in the following way: stealing is morally wrong; *x* is an instance of stealing; therefore, *x* is morally wrong. It is not necessarily the case that each individual step of reasoning is conscious in every instance of moral judging; some moral judgments are arrived at so quickly that they likely involve enthymematic reasoning of some sort, but it requires a good deal of practice to develop the appropriate moral expertise (see, for example, Gewirth, 1988; Hare, 1981; Herman, 1993; Wallace, 2008).[1] But, importantly, according to the deductive model of moral judgment, every moral judgment has its origin in some deduction, conscious or nonconscious, and the premises involved in that moral judgment are always consciously accessible, even if they are not consciously tokened in the process of moral judging.

The deductive model has long guided psychological research into moral judging, by focusing research on uncovering how the mature capacity for moral

[1] This view of moral judging is often attributed to Aristotle as well, based on his notion of the practical syllogism where practical action issues from the conclusion of deductive, syllogistic reasoning (see, *Nicomachean Ethics* 1147a26-32) (Aristotle, 1999). However, there are reasons to interpret Aristotle here not as claiming that the deductive model is an accurate psychological picture of moral judging, but that it rather provides a rational reconstruction (Hughes, 2003). Regardless of the best way to interpret Aristotle, it is important to distinguish the deductive model as a psychological picture of moral judging from a particular epistemic picture of what inferential relationships are required from a moral principle to a particular moral judgment in order for the moral judgment to be justified.

judging develops in terms of moral reasoning (Kohlberg, 1984; Piaget, 1932/1965).[2] Under the deductive model, to study the psychology of moral judging one simply needs to ask people to report their reasoning in response to a set of moral vignettes or dilemmas. These self-reports can then be used to characterize different forms of moral reasoning according to whether, for example, people reason from fine-grained moral principles or more general principles of morality, and how people weigh competing moral requirements.

There is an important assumption behind these methods of studying moral judging, and it is an assumption shared by those who accept the deductive model of moral judging: that the reasons people generally offer for their moral judgments are the actual basis of those judgments, and thus that verbal self-reports of one's reasons and reasoning is a reliable indicator of the actual processes of moral judging. Most people assume that the reasons people give for their moral judgments are the ones that actually enter into their moral reasoning, and are thus the actual basis of their moral judgments. Initially this seems like a fairly safe assumption, because introspectively it seems to most people that the reasons they give for their moral judgments are the actual basis of their moral judgments. Recent studies with respect to moral dumbfounding, however, challenge this assumption because they show that the reasons people offer for some of their moral judgments are not the actual basis of those judgments, and thus that verbal self-reports of one's reasons and reasoning is not a reliable indicator of the actual processes of moral judging.

[2] Though Piaget and Kohlberg's work has dominated psychological research into moral judgment, it is not without its share of critics. One notable criticism, by Gilligan, claims that Piaget's and Kohlberg's stage theories focus exclusively on justice, which ignores the different moral concerns of women that typically center on relationships and caring (Gilligan, 1982).

Moreover, there is a body of evidence showing that people lack access to the reasons underlying their evaluative judgments quite generally. For example, Nisbett and Wilson (1977) presented subjects with an array of identical objects, such as nylon pantyhose, and asked them to choose one. Most subjects showed a marked right-hand preference in selecting, but when questioned immediately afterward many subjects said the reason for their choice was, for example, that the one they chose was softer. Mood, too, has been shown to influence a wide range of evaluative judgments, including judgments of risk and blame, though people are generally unaware that their judgments have been influenced in this way (Finucane, Alhakami, Slovic, & Johnson, 2000; Schwarz, 2002). Winkielman et al. found that exposing individuals to a smiling face for 1/250th of a second increased how likable subjects rated ideographs that were presented to them immediately following the exposure (Winkielman, Zanjonc, & Schwarz, 1997). Importantly, subjects did not even know they had been exposed to a smiling face! Thus, it is not surprising that people may not always be aware of the reasons for their own moral judgments, but by challenging the assumption that the reasons people give for their moral judgments really are the basis of those judgments, moral dumbfounding studies challenge the supposed obviousness of the deductive model of moral judgment, and moreover, suggest that an alternative model of moral judgment may be necessary.

2. Moral Dumbfounding

The first challenge to the deductive model of moral judgment comes from studies on the phenomenon of moral dumbfounding. Consider the following case:

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night, they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love? (Haidt, 2001, p. 814).

According to research by Haidt et al. (Haidt, Bjorklund, & Murphy, 2000), most subjects judge that what Julie and Mark have done is morally impermissible, however, they are often unable to provide plausible reasons why incest, in this particular case, is wrong. When pressed to provide reasons why incest, in this case, is wrong, many subjects respond by saying that Mark and Julie will have deformed children, or that they will be emotionally scarred. When they are reminded that such problems could not occur in this case, they respond by saying something such as, "I know it's wrong, but I just can't come up with a reason why" (Haidt, et al., 2000).[3] In this sense, they are morally dumbfounded: they are unable to articulate any further plausible reasons for their moral judgment beyond the fact that it is a case of incest.

[3] There is some concern that Haidt's methodology, which involves face-to-face interviews might make it difficult for subjects to admit that their initial moral judgments were wrong, and thus "stick to their guns" even though they recognize that the reasons they cite do not support their moral judgments.

A number of studies suggest that moral dumbfounding is a rather common phenomenon. For example, Hauser and colleagues asked subjects to judge the Trolley case and some of its variants (Hauser, 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007). In the initial trolley case, a trolley is out of control and there are five people on the track ahead. There is no way to warn them, and there is no way for them to get out of the way in time. Throwing a switch will cause the trolley to go onto a sidetrack, where there is one person who will certainly be killed if the switch is thrown. Is it permissible to throw the switch?[4] Most people (around 90% according to Hauser (2006)) judge that it would be permissible to throw the switch, and when asked to give reasons people generally cite the difference in number as justifying throwing the switch. However, when the same respondents are asked to judge what seem to be structurally similar variants of the initial Trolley case, they judge the case very differently.

One such variant is the Footbridge case. Just as in the initial trolley case, there is a trolley that is out of control and threatens to kill five people on the track. A fat man is on a footbridge over the track, and if he is thrown over, his size is sufficient to stop the runaway trolley, thus saving the five people on the track. Is it permissible to throw the fat man over the footbridge? Most respondents judge that it would be impermissible to throw the fat man over the footbridge, even though the cases are

[4] The original trolley case is given by Phillipa Foot (Foot, 1967), and it is slightly different than the case presented here, which is Judith Jarvis Thomson's (Thomson, 1976). In Foot's original version, the question is whether the trolley *driver* should throw a switch that would cause the trolley to go onto a sidetrack. Foot argues that it is morally required for the driver to do so because if he does not, he or she kills five instead of one. Thomson introduces a bystander to the case because she argues that the bystander does not kill anyone if he or she does not throw the switch—the bystander can let five people die or kill one person. This subtle shift raises the question of whether there is any moral difference between killing and letting die, which is important to a number of contemporary moral debates.

numerically identical—one person dies to save five others. In most cases (70%), respondents are unable to articulate any coherent justification for judging the two cases differently (Hauser, 2006, p. 128). Notice that the problem is not that philosophers cannot find some principle that would explain the difference (such as the Principle of Double Effect), but that respondents make these different moral judgments, but are unable to offer coherent reasons for them: they are morally dumbfounded.[5,6]

Cushman et al. (Cushman, Young, & Hauser, 2006) found similar results using a somewhat different methodology. In Cushman et al.'s studies, subjects were asked to judge pairs of moral dilemmas that most people consistently judge differently (for example, in the first dilemma respondents would judge an action permissible, and in the second case impermissible). In each of the pairs of dilemmas, the divergent judgments could be explained by reference to one of three moral principles: the action principle (directly caused harm is morally worse than harm brought about by omission); the intention principle (intended harm is morally worse than harm brought about as a side-effect); and the contact principle (using physical contact to harm is morally worse than harm brought about without physical contact). What Cushman et al. found was that respondents reliably judged the pairs of cases in accordance with these three principles, but did not cite them in their justifications for their judgments (with the exception of the action principle). In many cases, the

[5] According to Hauser (2006), about 10% of subjects justified their different judgments by appeal to the Principle of Double Effect. However, Hauser's sample included highly educated scholars in many disciplines, and so the 10% who offered this justification may reflect the percent of participants with some familiarity with philosophy.

[6] Hauser's methodology also decreases the possibility that respondents are simply "sticking to their guns" with their initial judgments, and therefore confabulating, due to reputation effects. Hauser's study was conducted over the internet, and respondents never met the researchers, nor were respondents identifiable to the researchers.

justifications respondents offered were clearly implausible "prompted by the inability to justify the pattern of judgments" (p. 1086). In other words, they were morally dumbfounded.

The phenomenon of moral dumbfounding raises a serious problem for the deductive model of moral judgment, because it reveals that the reasons people offer for their moral judgments are not always the basis of those moral judgments. Instead, the reasons people offer for some of their moral judgments are obviously false, insufficient, or irrelevant, and even when this is pointed out, most people are unwilling to reconsider their moral judgment in any meaningful way.[7] This casts serious doubt on the supposed obviousness of the deductive model of moral judgment, because it reveals that the reasons people offer for their moral judgments are not related in the right way to the actual processes of moral judgment for people's self-reports of their reasons and reasoning to provide any reliable insight into the actual processes of moral judging. Moral dumbfounding alone does not show that the deductive model is wrong—it could be the case that the deductive reasoning processes are nonconscious—but it does call into question the evidential status of the reason's people give for their moral judgments as evidence of actual processes of moral judgment, and the deductive model of moral judgment in particular.

There is still one pertinent question with respect to moral dumbfounding: if the reasons people offer for their moral judgments are not the basis of those moral judgments, then what are they? Most likely, the reasons people offer in these cases are post hoc confabulations invented after the fact to rationalize their moral

[7] Haidt et al. report that only 16% of subjects revised their initial judgment with respect to Mark and Julie after researches pointed out that the case excludes the possibility that Mark and Julie could have children or that they would be emotionally scarred (Haidt, et al., 2000).

judgments. That is, it is likely that people make moral judgments first and that they invent reasons after the fact to support them. There is some evidence that post hoc confabulation is a fairly widespread feature of human psychology. A survey by Nisbett and Wilson found that subjects often provided confabulated reasons for their own actions—often citing reasons for their own behavior that could be shown by the experimenters to be false (Nisbett & Wilson, 1977). Gazzaniga found a similar (and more dramatic) pattern of post hoc justification of behavior in his research on so-called "split-brain" patients (Gazzaniga, 1995). "Split-brain" patients are those whose corpus callosum, which connects the two hemispheres of the brain has been severed or is absent. In Gazzaniga's study, he presented cards to a split-brain patient in such a way that simultaneous, yet different information was available to each side of the brain. When a card was flashed to the right brain saying, "Walk!" the patient got up and began to walk out of the testing area. When the researcher asked why, the patient responded that he was going to get a Coke. This reason was clearly confabulated, but to the patient it seemed like an ordinary introspective report of his own mental states.[8] Similarly, in the case of moral judgment it is likely that the reasons people offer for some of their moral judgments may be nothing more than post hoc confabulations, though they appear to the person to be ordinary introspective reports of their own mental states.

If self-reports are not a reliable guide to the actual processes of moral judging, then moral judging cannot be studied "head on" by asking people to report their reasons and reasoning to researchers. This has important methodological implications

[8] Gazzaniga argues from these findings that we really have no direct introspective access to the motives or causes of our judgments or behaviors, but rather, we are masters of interpreting our own behavior after the fact.

for the study of moral judging. To study the actual processes of moral judging researchers must devise ways of studying it obliquely, as it were, without relying on people's self-reports of their reasons for their moral judgments. Currently, three distinct research methods have been used to do this—behavioral studies, brain imaging studies, and research on psychopaths—and each has been used to argue that emotion, not consciously-accessible reasoning, plays an important causal role in moral judging.

3. Behavioral Studies

There is increasing evidence that subtle and not-so-subtle emotional manipulations reliably influence people's responses to moral vignettes. For example, a series of experiments by Schnall et al. found that feelings of disgust can increase the severity of subjects' responses to moral vignettes (Schnall, Haidt, Clore, & Jordan, 2008). In one experiment, Schnall and colleagues presented students with four vignettes after "fart spray" had been applied to a nearby trash bag. The presence and degree of "fart spray" applied increased the severity of subjects' responses to four moral vignettes. In a second experiment, Schnall and colleagues asked subjects to fill out a questionnaire, which, among other things, asked for their moral responses to three moral scenarios. One set of subjects were put at a filthy desk to fill out the questionnaire, and a second set of subjects were put at a clean desk. Those at the filthy desk reported more severe responses than those at the clean desk. In a third experiment Schnall and colleagues also found that viewing a disgusting video clip increased the severity of subjects' responses to moral vignettes compared with those who had been shown a neutral or sad video clip.

These experiments show that even subtle manipulations of people's feelings of disgust can increase the severity of their responses to moral vignettes. This is consistent with earlier experiments by Wheatley and Haidt, which showed that hypnotically induced disgust can make people's responses to moral vignettes more severe (Wheatley & Haidt, 2005). Wheatley and Haidt hypnotized people to "feel a sickening feeling" when they read either the word "often" or "take." They then presented subjects with vignettes, some of which included the hypnotic word. When the word was present, subjects responded more severely, that is, they indicated that the moral violations were more wrong than the control group. Not only did they find that disgust made people's responses to moral vignettes more severe, they found that it can also induce people to respond to a perfectly acceptable action as being more morally wrong when compared to the control group.[9]

Wheatley and Haidt included the hypnotic disgust word in the following mundane vignette, and asked subjects to indicate the degree of wrongness of Dan's actions:

> Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He [tries to take/often picks] topics that appeal to both professors and students in order to stimulate discussion (pg. 782).

[9] To quantify subjects' responses, Wheatley and Haidt had subjects make a slash mark along a 14cm (5.5in) line with one end labeled "not at all morally wrong" and the other end labeled "extremely morally wrong." These markings were then converted to a 100-point scale for analysis, with 100 being extremely morally wrong.

When subjects' hypnotic disgust word was present, they indicated that Dan's actions were more morally wrong than those whose disgust word was not present. Moreover, when subjects were asked why Dan's actions were morally wrong, they were hard-pressed to come up with any coherent justification. One subject wrote that "It just seems that he's up to something," and another wrote, "It just seems so weird and disgusting" (783).

While the forgoing studies exclusively focused on the influence of disgust in people's responses to moral vignettes, Valdesolo and colleagues designed an experimental protocol to test whether feeling happy affected people's responses to moral vignettes (Valdesolo & DeSteno, 2006). To test this, they presented half their subjects a clip from Saturday Night Live while the other half were presented an emotionally neutral video clip. Both groups were then presented the Trolley and Footbridge cases. What they found is that people who viewed a clip of Saturday Night Live before being presented with the Trolley case and the Footbridge case were much more likely to indicate that it is okay to push the fat man to his death in order to stop the trolley than those who did not (24% for those who watched the clip versus 8% in the emotionally neutral condition). So, it is not simply disgust that influences people's responses to moral vignettes, but being in a generally happier mood changes how one is disposed to respond to certain cases. And again, other research suggests that this is true of evaluative judgments in general, whether they be judgments of risk (Finucane, et al., 2000), blame (Schwarz, 2002), or the likability of ideographs (Winkielman, et al., 1997).

These behavioral studies show that emotions can have an appreciable influence on people's reactions to vignettes with moral and non-moral content, but they fall short of showing that emotions have an important causal role in the psychological processes of moral judging. The behavioral data only show that people's *behavior* in response to vignettes with moral and non-moral content (such as filling in a bubble on a questionnaire) is easily influenced by their emotions and mood, but this is not the same as showing that people's *moral judgments* are influenced by their emotions and mood. The reason is simple: the output that behavioral studies measure is overt behavior, not head-internal moral judgments. And this is an important difference, because it is unlikely that a person's moral judgments alone determine their overt behavior. How people actually behave is ifluenced by complex interactions among their moral judgments, emotions, mood, and other social and experimental cues. What behavioral studies actually study is the output of these complex interactions, not a person's moral judgment. Thus, behavioral studies alone are not sufficient to show that emotions have an appreciable influence on people's moral judgments. Other research, however, promises to show that emotions do have an appreciable influence on people's moral judgments, and moreover, promises to show how.

4. Brain Imaging Studies

Given the limitations of behavioral studies, it would be better to just look at what was going on in someone's mind when that person produced a moral judgment to see how various psychological processes contribute to its production. The one serious limitation here is that it is simply not possible to look into a person's mind.

But researchers are able to look at brains with fMRI machines,[10] and one recent trend in empirical research with respect to moral judging is to scan the brains of individuals while they are producing moral judgments. Such scans can provide, in some sense, a snapshot of the "moral brain" in action, allowing us to see which brain areas are active, and how, if it all, the "moral brain" looks different from the brain that is engaged in non-moral forms of cognition. This research at least promises to provide some useful insights into the neurological correlates of moral judging, and assuming some degree of correspondence between brain areas and psychological processes,[11] it could provide some insights into the causal role of emotions in moral judging (Greene & Haidt, 2002).

To get at the difference between the "moral brain" and the brain engaged in non-moral forms of cognition, Moll and colleagues ran an experiment where they placed ten subjects in an fMRI machine and read a series sentences to them while scanning their brains (Moll, Eslinger, & de Oliviera-Souza, 2001). Half of the sentences read to the subjects had moral content, such as "Old people are useless," or "They hung an innocent," while the other half were factual sentences, such as "Stones are made of water," or "Walking is good for your health." Subjects were instructed to judge whether the sentence read to them was right or wrong. (The words "right" and

[10] fMRI stands for "functional magnetic resonance imaging," and it measures the amount of blood flow in localized brain areas. Increased blood flow is a proxy measure for increased neural activity, indicating that a particular brain area is currently active. More detailed information is available at: www.fmri.org/fmri.html, maintained by the Columbia University Medical Center Program for Imaging and Cognitive Sciences.

[11] It is important to distinguish the brain from the mind: the brain referring to specific neurological and physical processes whereas the mind refers to specific mental processes. How the mind and brain relate to one another is, of course, subject to intense philosophical dispute. Chomsky attempts to sidestep this dispute by arguing that the mind and brain refer to the same object at different levels of abstraction, and so he uses the terminology mind/brain (Chomsky, 1980). This is an attractive position, but I think too many researchers ignore the difference between the mind and the brain altogether, and so I will maintain conventional usage, and use the term mind when referring to psychological/mental processes, and brain when referring to neurological/physical processes.

"wrong" were used intentionally because they cover both senses of being morally right and wrong and being factually correct or incorrect).

The fMRI scans revealed that there is, indeed, a difference between the "moral brain" and the brain engaged in non-moral forms of cognition and evaluation. When evaluating sentences with moral content, the subjects showed increased activity in their Frontal Polar Cortex (FPC) and right anterior temporal cortex; an increase that was not observed when subjects evaluated factual sentences. This result is significant because the areas of the brain that showed increased activity while evaluating sentences with moral content are associated with specific types of emotional processing, in particular, empathy and attention to one's own subjective emotional states. Thus, one difference between the "moral brain" and the brain involved in non-moral forms of cognition is that the "moral brain" recruits certain emotion centers, which suggests, again, that emotions have some causal role in the production of moral judgments.

Following up on these initial findings, Moll and colleagues employed a similar research design where they performed fMRI scans on seven subjects while displaying visual images (Moll et al., 2002). Some of the visual images contained morally salient content, such as physical assaults, abandoned children, and war scenes, while others contained unpleasant, pleasant, and neutral content. Moll and colleagues found that the visual images with morally salient content more reliably and differentially activated the right medial orbital frontal cortex (OFC), the medial frontal gyrus (MedFG), and the right posterior superior temporal sulcus (STS), as well as the amygdala, insula, thalamus, and upper midbrain; as did unpleasant images.

Importantly, these brain areas are associated with the processing of social and emotional events. Again, providing some evidence that emotions figure causally in the production of moral judgments.

While Moll and colleagues were concerned with seeing how the "moral brain" differs from the brain engaged in non-moral cognition, Greene and colleagues used fMRI scans to determine whether the differential activation of emotion centers in the brain could explain the difference in people's responses with respect to the Trolley case and the Footbridge case, among others (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Recall that most people indicate that it permissible to throw the switch to divert a runaway trolley that will kill one person but save five, whereas most people indicate that it is impermissible to throw the fat man in front of the trolley to save five people (Hauser, 2006; Hauser, et al., 2007; Mikhail, 2007, 2009; Petrinovich, O'Neill, & Jorgensen, 1993). Greene and colleagues hypothesized that a personal moral dilemma, involving up-close-and-personal contact with another person, would activate emotion centers in the brain causing the person to indicate that the action is wrong, as in the Footbridge case. Impersonal moral dilemmas, on the other hand, which do not involve up-close-and-personal contact, they hypothesized would not activate these emotion centers, leading to a more calculated utilitarian response, such as in the Trolley case.[12]

[12] There are some serious problems in the operationalization of the concepts moral-personal and moral-impersonal dilemmas. A personal dilemma as one that satisfies all three of the following criteria, otherwise it is impersonal: (1) it could reasonably be expected to lead to serious bodily harm; (2) to a particular person; and (3) the harm is not the result of "deflecting" a harm (Greene et al., 2001 pg. 2107 *n*. 9). The first difficulty with this categorization is that it is entirely *ad hoc*, designed specifically to capture the difference between the Trolley and the Footbridge cases. Secondly, it does not neatly divide the universe of possible moral dilemmas in ways that capture our intuitive notion of up-close-and-personal violations (Mikhail, 2009). Greene and colleagues admit that this distinction is a working one when they write, "This personal-impersonal distinction has proven useful in generating the present

To test this hypothesis, Greene and colleagues posed sixty practical dilemmas to subjects while scanning their brains in an fMRI machine. Among these practical dilemmas were moral-personal dilemmas, including the Footbridge case, and moral-impersonal dilemmas, including the Trolley case. Their results confirmed their hypothesis: moral-personal dilemmas differentially activated emotion centers of the brain, including medial portions of Broadman's area, medial frontal gyrus (MedFG), posterior cingulate gyrus, and the angular gyrus than either moral-impersonal dilemmas or non-moral practical dilemmas. These findings were confirmed by a subsequent follow up study by Greene and colleagues, which also showed that subjects had longer reaction times to moral-personal dilemmas than moral-impersonal dilemmas (Greene, Nystrom, Engell, Darley, & Cohen, 2004).

fMRI experiments reveal that there is a difference between the "moral brain" and the brain engaged in non-moral forms of cognition, namely, that the "moral brain" recruits brain centers associated with the processing of emotions, particularly when the scene of evaluation involves a moral-personal dilemma. This is a significant finding, if not a surprising one, but there is no way, using these brain scans, to distinguish among the various causal relationships that might exist between the emotions and a moral judgment, because fMRI scans are not sufficiently fine-grained. As Huebner and colleagues observe:

---

results, but it is by no means definitive. We view this distinction as a useful 'first cut'" (Greene, et al., 2001, p. 2107). In later works, however, Greene attempts to justify this distinction, as is, on evolutionary grounds (Greene, 2005, 2008; Greene & Haidt, 2002), and draws stronger and stronger conclusions from the data derived from it. In his (2008), for example, he argues that we must reconceptualize the entire history of moral philosophy based on this distinction and nine fMRI scans—all on the basis of an approximate and "first cut" distinction that has obvious conceptual problems.

> the activity of emotional circuits provides only correlational data, showing that emotions are associated with moral judgments. Such data (on their own) can never be used to infer causality, and because of the poor temporal resolution of neuroimaging, cannot be used to assess when emotions have a role or whether they are constitutive of moral concepts" (Huebner, Dwyer, & Hauser, 2009).

Neuroimaging is a blunt tool that does not allow for fine-grained discriminations among possible ways emotions might figure causally in moral judgment, or whether it is simply correlated with some instances of moral judgment, and so findings from fMRI machines cannot show how the emotions figure causally in moral judging.

5. Psychopathy and "Acquired Sociopathy"

The promise of brain scans was that they would allow us to see the "moral brain" in action in such a way that we could tease apart the causal contributions of various psychological processes in the production of a moral judgment. Unfortunately, brain scans are not sufficiently fine-grained enough to live up to that promise. But, another way to see what the "moral brain" or "moral mind" is doing is to see whether there are any neurological and/or psychological differences between those who have a normal capacity for moral judging and those who have some deficit in moral judging. This would allow us one way of seeing what neurological or psychological processes are that make moral judging possible, and could possibly

shed some light on the causal role of emotions in moral judging. It is this possibility that makes research on psychopaths so interesting.

According to Cleckley, who wrote the first major work on psychopathy, psychopaths, as a class, are often petty, and sometimes violent, criminals (Cleckley, 1964). They rarely, if ever, develop any meaningful plans for their lives, and when they do, lack the motivation to carry them out. Their actions are often impulsive and irrational, in the sense of serving no interest of their own and in being almost wholly inexplicable. Psychopaths regularly flout social norms and engage in immoral behavior, without experiencing any guilt or remorse for doing so (though as Cleckley observes, they are often quite good at *expressing* guilt and remorse, though further questioning reveals that they do not actually *experience* these emotions). Even though psychopaths often engage in personally and socially destructive behavior, they are generally quite charming and master manipulators of those around them, including their family and friends, judges and lawyers and well-meaning psychiatrists, and so often avoid social and legal sanction.

Alongside these observational findings, further research by Blair has shown that psychopaths are far less likely to treat moral violations differently than conventional transgressions; that psychopaths are more likely to give conventional justifications for moral violations; and that psychopaths are less likely to consider harm or pain in their justifications of moral rules (Blair, 1995; Blair, Jones, Clark, & Smith, 1995; Blair, Monson, & Frederickson, 2001). Moreover, children with psychopathic tendencies are more likely to judge moral rules as authority contingent than children in the control group (R. J. R. Blair, 1997). For the psychopath, then,

there is no real difference between the judgment that it is wrong to hurt people for fun and the judgment that it is wrong to chew gum in class.

What makes psychopaths so interesting for the study of moral judging is that even though they are strikingly different in their behavior from ordinary individuals, their ability to reason abstractly is entirely intact. Psychopaths score generally well on IQ tests, and show no discernible difference in intelligence from the normal population (Damasio, 1994). Moreover, there is no discernible difference between psychopaths and normal individuals with respect to their ability to reason, even morally. As Cleckley writes:

> Despite the extraordinarily poor judgment demonstrated in behavior, in the actual living of his life, the psychopath characteristically demonstrates unimpaired (sometimes excellent) judgment in appraising theoretical situations. In complex matters of judgment involving ethical, emotional, and other evaluational factors, in contrast with matters requiring only (or chiefly) intellectual reasoning ability, he also shows no evidence of defect. So long as the test is verbal or otherwise abstract, so long as he is not a direct participant, he shows that he knows his way about. He can offer wise decisions not only for others in life situations, but also for himself so long as he is asked what to do (or should do, or is going to do). When the test of action comes to him we soon find ample evidence of his deficiency (pp. 367-8).

Whatever causes psychopaths to engage in immoral and anti-social behaviors it cannot be attributed to any deficit in their ability to reason abstractly and deductively, even about moral questions. And what is wrong with psychopaths is that they lack the ability to process and experience emotions, in particular, they lack the capacity to experience empathy (Cleckley, 1964; Damasio, 1994; Kiehl, 2008). Blair found that children with psychopathic tendencies were less responsive to distress cues in other individuals, as measured by their skin conductance responses (Blair, 1999). fMRI investigations reveal that psychopaths have less activity in the emotion centers of the brain than normal individuals when subjected to mild pain (Birbaumer et al., 2005) and when shown pictures of threatening faces and severely injured people (Muller et al., 2003). Cleckley describes psychopaths as having flat affect, and notes that they are generally unresponsive to the emotional distress of others.

Some theorists have argued that these findings indicate that one important difference between the properly functioning "moral mind" and the dysfunctional psychopathic mind is the ability to experience emotions, in particular, empathy, and that this suggests that emotions are, in some way, *necessary* for moral judging (Blair, 2009; Nichols, 2004; Prinz, 2007). This necessity claim, however, is quite strong, and goes well beyond the most plausible explanation of what is wrong with the psychopath, namely, a disconnect between their moral judgments and moral action. This is a point I shall return to later, but before I do, it is important to get clear on what this necessity claim amounts to. There are two ways to understand the claim that emotions are necessary for moral judging. On the one hand, it could be that the

tokening of an emotion is necessary for the tokening of a moral judgment. On the other hand, it could be that emotions are necessary for the proper development of our capacities for moral judgment. The data with respect to psychopathy alone do not distinguish between these two ways emotions might be necessary. However, there is some interesting findings on those with so-called "acquired sociopathy" that can help decide between these two possibilities.

"Acquired sociopathy" is a condition where people with normally functioning moral capacities develop similar patterns of behavior as "natural born" psychopaths after suffering specific head trauma or brain lesions to the ventromedial prefrontal cortex (VMPFC) (Damasio, 1994).[13] The specific brain areas whose damage leads to "acquired sociopathy" are those associated with having certain emotions, and individuals with acquired sociopathy have greatly reduced social emotions, such as empathy and guilt (Damasio, 1994; Kiehl, 2008). And, just like "natural born" psychopaths, those with "acquired sociopathy" generally have intact capacities for deliberation, including moral deliberation (Damasio, 1994; Saver & Damasio, 1991).

If emotions were only necessary for the proper development of the capacity for moral judging, then damage to the emotion centers later in life should leave the capacity unchanged. However, if the tokening of emotions in necessary for the proper functioning of the capacity for moral judging, then damage to the emotion centers should bring with it a consequent dysfunction in one's capacity for moral judging. It is this latter outcome that we find: take a normal person and remove his or her ability

[13] Damasio coined the term "acquired sociopathy" but it is not recognized by the American Psychological Association as a distinct disorder, and so I keep it in quotes throughout. There is an important distinction between psychopathy and "acquired sociopathy" in terms of its causal etiology, but not in terms of clinical diagnosis.

to experience certain emotions, and you end up with someone very much like a psychopath.

This finding has led some philosophers to argue that the ability to token emotions is necessary for the proper functioning of the capacity for moral judging, and thus that tokening an emotion is *necessary* for moral judgment (Nichols, 2002, 2004; Prinz, 2007).[14] Their reasoning is straightforward enough: psychopaths show no disability in their ability to reason deductively, even with respect to moral questions, however, there is something quite wrong in the way that psychopaths behave—they are not motivated by their moral judgments, nor do they find them any more compelling than an instruction from a school teacher to raise one's hand before asking a question. This disconnect between reasoning ability and moral behavior provides strong empirical evidence against the deductive model of moral judging; if moral judgments are the conclusions of deliberation, and psychopaths can do that just fine, and one assumes that moral judgments necessarily motivate, then whatever we ordinarily think of as moral judgments, it cannot simply be the conclusion of moral reasoning. Moreover, what is wrong with psychopaths is their inability to process and experience emotions, in particular, empathy. Since psychopaths cannot produce what we ordinarily consider moral judgments, and their only deficit is in processing and experiencing certain emotions, then these emotions must be necessary for producing moral judgments.

[14] Nichols and Prinz are hardly alone among philosophers in arguing that psychopaths do not make genuine moral judgments. The view is quite popular among defenders of moral judgment internalism, who hold that moral judgments, in some sense, necessarily motivate (Cholbi, 2006a, 2006b; Hare, 1952; Smith, 1994).

The strength of this necessity argument, however, hangs on accepting the metaethical claim that moral judgments necessarily motivate. But this claim is widely disputed (see, for example, Brink, 1989; Roskies, 2003, 2006; Sadler, 2003; Svavarsdottir, 1999, for arguments against moral judgment internalism, as this view is known). A competing hypothesis is that the capacity for moral judging is undamaged in psychopathy, but that the normal emotions and motivations that accompany moral judgments are absent. On this view, what is wrong with psychopaths is that they lack moral *agency*, but not the capacity for moral *judging* (Kennett, 2006; Roskies, 2003, 2006). For present purposes, however, it is important to outline how contemporary moral psychologists have interpreted the empirical data surveyed in this chapter.

6. New Wave Sentimentalism

Taken as a whole, recent empirical work casts serious doubt on the empirical plausibility of the deductive model of moral judgment. It does not show that the deductive model is wrong by any means, and there are still plenty of philosophers who defend the deductive model of moral judgment in light of these empirical findings (Fine, 2006; Horgan & Timmons, 2007; Jones, 2003, 2006; Kennett, 2006; Kennett & Fine, 2009). But, as in most empirical disputes, one rarely lands a knockout blow. Rather, the test for any empirical model is how well it accounts for the available evidence, whether it is consistent with other known facts, and how simple and elegant it is compared with other alternatives. And there are many theorists who have taken the findings from moral dumbfounding, behavioral studies, brain imaging studies, and psychopathy to develop sentimentalist accounts of moral judging that largely discard the role of reasoning in moral judging. I call these views

new wave sentimentalist views to differentiate them from more traditional sentimentalist accounts of moral judging which focus on supposed analytic connections between moral concepts and emotions (such as, for example, Ayer, 1952; D'Arms & Jacobson, 2000b; Gibbard, 1990; Stevenson, 1937). New wave sentimentalism is primarily motivated by empirical findings, and cast itself as an empirical (as opposed to analytic) approach to studying moral judging. Moreover, whereas traditional sentimentalist views maintain an important role for moral reasoning in moral judging, new wave sentimentalist models of moral judging do not. The two most well developed and influential new wave sentimentalist models of moral judging are the Social Intuitionist Model, and Constructive Sentimentalism.

Before outlining those views, I should note that there are a range of empirically-motivated models of moral judging that are not sentimentalist, and flesh out the nonconscious processes of moral judgment in terms of a Universal Moral Grammar (Dwyer, 2006, 2009; Hauser, 2006; Mikhail, 2007, 2009), heuristics and biases (Gigerenzer, 2008; Sunstein, 2005), or some other nonconscious reasoning processes (Greene, 2008; Sripada & Stich, 2006). I am not going to spend time outlining these views, because my central concern over the next few chapters is with two specific claims of new wave sentimentalists that have gained considerable traction among contemporary moral psychologists, and as a result, have set the terms of the contemporary debate with respect to moral judging and moral psychology more generally. The first claim is that the fact that nonconscious processes are causally important in moral judging requires abandoning a central causal role for consciously

accessible reasoning in moral judging.[15] The second claim is that metaethical conclusions can be "read off" of psychological claims; that is, that some metaethical conclusions follow *directly* from psychological claims. One of the central contentions of this dissertation is that neither of these claims are warranted, and thus that the contemporary debate with respect to moral judging and moral psychology has been framed in the wrong way. It is because new wave sentimentalist views have been so influential in framing the terms of the contemporary debate that I focus on their models of moral judging.

6.1 Social Intuitionist Model

The Social Intuitionist Model is by far the most influential new wave sentimentalist model of moral judging. Haidt gives the original formulation of the model in his (2001), but in subsequent years Haidt and colleagues have modified the model to incorporate other findings and address certain criticism (Haidt & Joseph, 2007; Haidt & Joseph, 2004; Wheatley & Haidt, 2005).[16] Haidt and Bjorklund (Haidt & Bjorklund, 2008a, 2008b) give the most recent and complete formulation of the Social Intuitionist Model, and I will treat those papers as the current mature view.

Social Intuitionists take themselves, in the first instance, to be offering an explanation of the capacity for moral judging, but there are some problems in understanding precisely what Social Intuitionists take the causal processes and

[15] Among theorists who explicitly endorse this claim and attribute it to the Social Intuitionist Model are Gigerenzer (2008), Hauser (2006), and Mikhail (2007, 2009).
[16] See (Narvaez, 2008; Pizarro & Bloom, 2003; Saltzstein & Kasachkoff, 2004) for criticism of the view.

informational processes of moral judging to be. According to the Social Intuitionist Model, moral judgments are largely driven by moral intuitions, which they define as:

> the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like-dislike, good-bad) about the character or action of a person, without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion (2008a, p. 188).

In the ordinary case, Social Intuitionists claim that these moral intuitions "lead directly" (pg. 188) to a moral judgment, which on this view is "the conscious experience of blame or praise, including a *belief* in the rightness or wrongness of the act" (2008a, p. 188, emphasis in original). But again, the precise informational and causal processes of the capacity for moral judging are not spelled out by Social Intuitionists. They claim that there is a "tight connection between flashes of intuition and conscious moral judgments" (pg. 188), but a tight connection is not necessarily a causal one. This is an important problem, because it is difficult to understand what, precisely, Social Intuitionists take the role of emotions to be in moral judging. And as Dwyer notes, failing to spell out the nature of the link between intuition and judgment renders the supposed role of emotions in moral judging mysterious, unless Social Intuitionists simply want to stipulate that "moral judgments are moral intuitions made conscious" (Dwyer, 2009, p. 277).

This is a deep problem for the Social Intuitionist Model, but in order to develop the problems I want to address, I will set it aside for now and give what I

take to be the most charitable interpretation of their claims. On this interpretation, the Social Intuitionist Model claims that certain emotional reactions (moral intuitions) are a sufficient cause of a person's moral belief that some action is right or wrong, and this moral belief, attached to the emotion, is the moral judgment. Moreover, the initial emotional reactions have no basis in consciously accessible reasoning, and thus, according to Social Intuitionists, when people are pressed to provide reasons for their moral judgments, they confabulate reasons after the fact. As they say, people are like lawyers who try to provide the best defense and argument possible for their position, but their goal is not to discover the moral truth, as it were, but simply to defend their emotionally caused moral judgments.

That is the intuitionist part of the Social Intuitionist Model. The social part, however, is just as important. According to Social Intuitionist Model, moral judging is not fully explained by one's private cognitions, because how one's intuitions and judgments affect the intuitions and judgments of others, and vice versa, is an important causal mechanism of moral judging. The post hoc reasons a person may offer for her judgment have the potential to influence the moral judgment of others, either through direct social pressure, or because the reasons force her to view the situation in a different light, causing her to have different moral intuitions and therefore leading to a new moral judgment. Moreover, post hoc reasoning can also cause a person to change his or her own moral judgment, so long as those post hoc reasons trigger in that person a new and different moral intuition.

Reasoning can play some role in moral judgment beyond simply producing confabulated reasons. According to the Social Intuitionist Model, reasoning alone can

sometimes lead to a moral judgment through the "sheer force of logic" (Haidt & Bjorklund, 2008a, p. 193) and can sometimes directly override a conflicting moral judgment, if the moral intuition that caused it is weak. Social Intuitionists hypothesize that this is very rare, however, accounting for less than 5% of moral judgments (Haidt & Bjorklund, 2008a, p. 212). Figure 2.1 gives the current formulation of the Social Intuitionist Model.

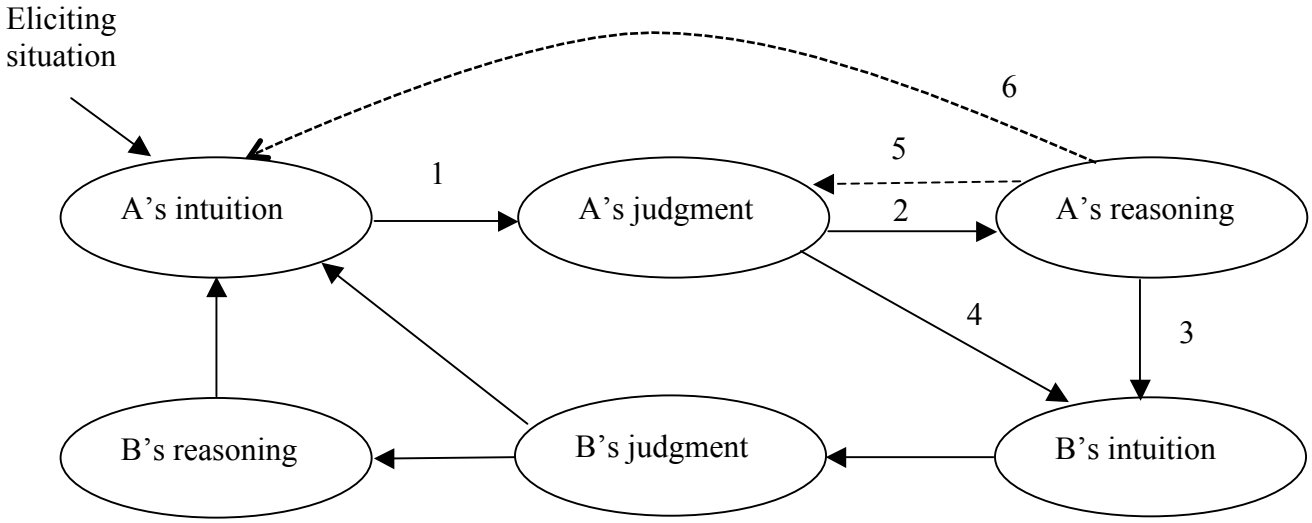

Figure 2.1 outlines the Social Intuitionist Model. Reprinted from Haidt and Bjorklund (2008a, p. 187) with corrections (the original diagram has B's arrows in the wrong direction). In the figure (1) is the intuitive judgment link; (2) is the post-hoc reasoning link; (3) is the reasoned persuasion link; (4) is the social persuasion link; (5) is the reasoned judgment link; and (6) is the private reflection link.

The Social Intuitionist's view of the role of reasoning in moral judgment has gained considerable traction, even among those who reject their characterization of the nonconscious processes of moral judging. For example, Hauser, who argues for a Universal Moral Grammar writes, "Conscious moral reasoning often plays no role in

our moral judgments, and in many cases reflects a post-hoc justification or rationalization of previously held biases or beliefs" (Hauser, 2006, p. 25). And Gigerenzer considers it an upshot of his view that intuitive moral judgments are the result of heuristic processes that it, "fits well with the social intuitionist view of moral judgment, where rationalization is ex post facto rather than the cause of the decision" (Gigerenzer, 2008, p. 15). The point from all these empirical moral psychologists is the same: moral reasoning of the sort envisioned by the deductive model of moral judgment is nearly impossible and very rare—moral reasoning only *appears* to be the conscious consideration of various reasons and principles—but the real psychological story is quite different. Moral reasoning is post hoc justification of already arrived at moral judgments, but the real causes of moral judgment are nonconscious processes.

6.2 Constructive Sentimentalism

Another new wave sentimentalist account of moral judging is given by Prinz (Prinz, 2007), who labels his model Constructive Sentimentalism. The central motivation for Constructive Sentimentalism is not psychological, but metaethical. According to Constructive Sentimentalism, the best interpretation of the empirical data is that emotions are *necessary and sufficient* for a moral judgment because emotions are constituents of moral concepts. For example, according to Constructive Sentimentalists,[17] the concept WRONG is "a detector for the property of wrongness that comprises a sentiment that disposes its possessor to experience emotions in the

[17] Prinz is the only Constructive Sentimentalist on record, but I use the plural "Constructive Sentimentalists" because it is a less wieldy locution.

disapprobation range" (Prinz, 2007, p. 94). Similar analyses apply, *mutatis mutandis,* to other moral concepts such as RIGHT, GOOD, and BAD.

This conceptual claim, however, is rather problematic. Conceptual analyses generally attempt to specify what a concept *means*, that is, what someone means when they use a term that refers to a concept. But it is far from obvious how, for example, the statement "murder is wrong" just means that "murder is a detector for the property of wrongness that comprises a sentiment that disposes its possessor to experience emotions in the disapprobation range." This conceptual analysis turns a perfectly intelligible English sentence into one that, if not unintelligible, is certainly unclear. This is a serious problem for the Constructive Sentimentalist's metaethical claim, but what is important for present purposes is the causal story of the processes of moral judging that Constructive Sentimentalists argue supports this conceptual analysis.

Prinz uses the example of pickpocketing to explain the Constructive Sentimentalist view of the processes of moral judging, and we can use that example here (pg. 96). First, a person perceives an event of pickpocketing and categorizes that event as falling under some concept. In the case of pickpocketing, the person categorizes the event as falling under the concept of STEALING. Importantly, this categorization step does not constitute the moral judgment, because, on the Constructive Sentimentalist view, morally relevant concepts, such as STEALING, do not have any moral content, such as *wrongful* taking—they are purely descriptive.[18] If the person has a rule against stealing, though, then classifying the perceived event as stealing activates that moral rule in long-term memory. Importantly, according to

[18] I shall challenge this view of morally relevant concepts in the next chapter (pp. 72-75).

Constructive Sentimentalists, a moral rule is a sentiment toward a concept, which disposes a person to produce an emotion. For example, if someone has a rule that stealing is wrong, that rule is a sentiment towards stealing that disposes the person to produce an emotion in the disapprobation range (anger, guilt, shame, etc). Once a moral rule has been activated, contextual factors determine which emotion is produced. For example, if someone else has transgressed the rule, the person will feel anger, whereas if the person broke the rule, he will feel guilty. Lastly, the emotion is bound to the representation of the perceived event that elicited the emotion. So, in this example, the person will feel anger towards the act of pickpocketing. Figure 2 outlines the process of moral judgment according to Constructive Sentimentalists.

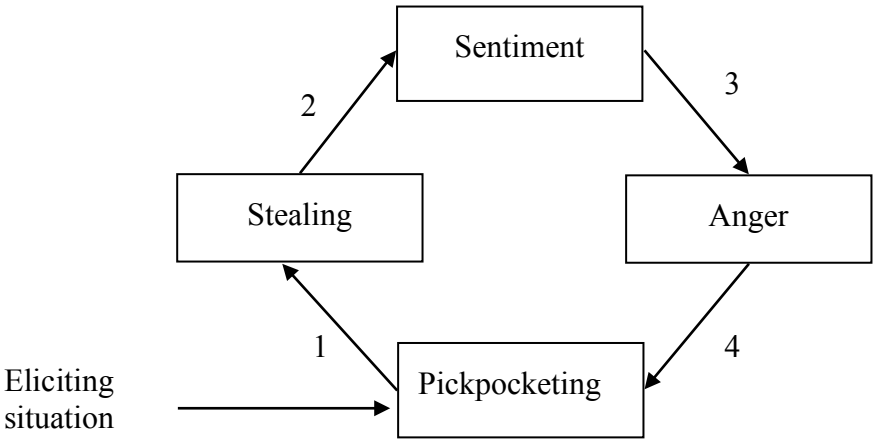

Figure 2.2 Information-processing stages that lead to a moral judgment. (1) A perceived event is categorized; (2) a rule is retrieved from memory, which activates a sentiment; (3) the sentiment elicits an emotion in a contextually-sensitive way; (4) the emotion is associated with the perceived action. (Adapted from Prinz, 2007 p. 97).

According to Constructive Sentimentalists, emotions are necessary and sufficient for moral judging, so what role can Constructive Sentimentalists possibly allow for moral reasoning? They maintain that reasoning has two possible roles in moral judging. First, just as Social Intuitionists have it, one role of reasoning in moral judging is to provide post hoc justifications of one's already arrived at moral judgments. A second role for reasoning in moral judging, according to Constructive Sentimentalists, is that a person can reason about whether a given event falls under a particular concept towards which a person has a sentiment, that is, a moral rule. This reasoning can be intra- and interpersonal. In the intrapersonal case, for example, a person may wonder if a given foreign policy is morally permissible, and she may have to deliberate for hours to determine whether it falls under any of her concepts towards which she has a sentiment. If she finally determines that it does, it will immediately elicit a moral judgment (pg. 114). In the interpersonal case, moral reasoning can involve two or more people providing each other reasons for thinking that a certain event falls under a given concept towards which they both have a sentiment (pg. 125). For example, the debate over abortion often involves arguing over whether abortion falls under the concept murder or not. Assuming that all participants have a sentiment against murder, this sort of reason giving constitutes moral reasoning, according to Constructive Sentimentalists.

Prinz argues that the role of reasoning in moral judgment is much more expansive in the Constructive Sentimentalist model than in the Social Intuitionists Model. He writes:

> I agree with Haidt that reasoning often plays this role [post hoc justification], but reasoning can also play an important role in determining whether a certain event falls under a category about which we have a moral sentiment. We often have to reason to determine whether something is a case of discrimination, for example. Haidt's model allows for this but he doesn't emphasize it. That's an important oversight, because a lot of moral debate may involve rational disagreements about how to categorize things. But once we've categorized something as a case of discrimination, reasoning stops. There is an immediate emotional reaction (pg. 124).

***

Though Social Intuitionists and Constructive Sentimentalists claim that moral reasoning has some role in moral judging, they both agree that it has a much more contracted role than supposed by the deductive model of moral judgment. In this way, these models can be seen as offering an important corrective to the deductive model, which they argue gives too prominent a role to reasoning in moral judging and ignores the important causal contribution of the emotions. Now, it is quite possible that the deductive model of moral judgment gives too much of a role to moral reasoning, but the Social Intuitionist and Constructive Sentimentalist models of moral judgment threaten to give too little a role to moral reasoning because they exclude some roles for reasoning in moral judging that are thought to be fairly typical, including reasoning to determine whether one's own moral judgments are coherent

and consistent; whether one's moral judgments cohere with one's other moral attitudes and commitments; and reasoning to determine whether one's moral judgments are appropriate.

In this respect, these models of moral judging raise two important questions: (1) does recognizing that emotions have a central causal role in moral judging *require* giving up a central causal role for distinctively moral reasoning; and (2) can such a view of moral judging, which gives up a central causal role for distinctively moral reasoning, that is, moral reasoning with distinctively moral content, satisfy the constraint that a model of moral judgment must be consistent with other important features of our moral psychology? Over the next three chapters, I shall argue that the answer to both of these questions is "no." However, an important methodological note is in order before proceeding: as research on moral dumbfounding makes clear, it is not sufficient in answering this question to assert that reasoning does have a central causal role in moral judging based simply on one's own introspective experiences of moral reasoning. Introspection is not a reliable guide here; at least not in such a way that introspection alone can be sufficient to show that moral reasoning does have these roles. Rather, the way to proceed is from the side, by seeing whether the roles for moral reasoning envisioned by Social Intuitionists and Constructive Sentimentalists are consistent with further robust phenomena of our psychology. This is the second condition of adequacy on any model of moral judging, and it is here, I shall argue, both Social Intuitionism and Constructive Sentimentalism come up short.

7. Conclusion

Recent empirical findings raise serious problems for the deductive model of moral judgment, challenging some fairly entrenched assumptions with respect to the role of moral reasoning in moral judging. Based on the empirical research, new wave sentimentalist models of moral judging maintain that emotions have a central causal role in moral judgment, and greatly minimize the role of moral reasoning in moral judging. Going forward, the question I shall raise is whether new wave sentimentalists are right in maintaining that recognizing an important causal role for the emotions in moral judging requires minimizing the role of moral reasoning. I shall argue that it does not.

Chapter 3: Moral Reasoning & Moral Change

The question I want to turn to in this chapter is whether recognizing that the emotions play an important causal role in moral judging requires setting aside the possibility that moral reasoning can also play an important causal role in moral judging. Answering that question, however, requires saying something about what moral reasoning consists in, that is, what processes constitute the capacity for moral reasoning. Both Social Intuitionists and Constructive Sentimentalists argue that the fact that emotions play an important causal role in moral judging requires abandoning the deductive model of moral judgment, along with its account of moral reasoning as deduction from first principles of morality. However, neither Social Intuitionists nor Constructive Sentimentalists argue that moral reasoning is therefore causally inert or ineffective in moral judgment; rather, they each argue that moral reasoning consists in some other set of processes that fits naturally within their respective sentimentalist models of moral judging. The aim of this chapter is to argue that Social Intuitionist's and Constructive Sentimentalist's respective accounts of what moral reasoning consists in are not consistent with broader features of our moral psychology, in particular, moral change. Both of these accounts come with some significant costs to explanatory adequacy, and thus neither satisfy the second constraint on any model of moral judging: that it be consistent with the broader facts of our moral psychology. I conclude by offering my own naturalized account of what moral reasoning consists in that avoids these problems.

1. The Central Issue

Many theorists have criticized the Social Intuitionist and Constructive Sentimentalist models of moral judging by objecting to their respective accounts of moral reasoning, specifically, by arguing that these models of moral judging problematically minimize the role of reasoning in moral judging (Bloom, 2010; Jones, 2006; Kennett & Fine, 2009). The central thrust of these criticisms is that moral reasoning occurs *more often* and that it can have more of a *direct causal effect* on moral judging than Social Intuitionists or Constructive Sentimentalists allow, and thus that these models of moral judging are wrong, or at least incomplete. Kennett and Fine (2009), for example, criticize the Social Intuitionist Model by arguing that moral reasoning often involves overriding a moral intuition, and thus that Social Intuitionists Model wrongly minimizes how often moral reasoning has a direct causal effect on moral judging. Bloom similarly criticizes the Social Intuitionist Model by asserting that the model involves the "wholesale rejection of reasoning" (2010, p. 490). Jones (2006) argues that the Constructive Sentimentalist model does not allow that moral reasoning can have a direct causal effect on moral judging, and thus the model entails that moral judgments are not reasons-responsive, and thus that ordinary people are not moral agents.

These criticisms, however, fall wide of the mark. Both Social Intuitionists and Constructive Sentimentalists allow that moral reasoning can occur quite often in some contexts, and that when it does, it can have a direct causal effect on moral judging. Prinz claims that the Constructive Sentimentalist model leaves "plenty of room for rational debate. We often need to use reason to demonstrate that an action falls under a moral category...[which then causes] an immediate emotional reaction," that is, a

moral judgment (Prinz, 2007, p. 124). And Social Intuitionists claim that, "The core of the model gives moral reasoning a causal role in moral judgment, but only when reasoning runs through other people" (Haidt & Bjorklund, 2008a, p. 193). Both Social Intuitionist and Constructive Sentimentalists hold that moral reasoning can occur fairly often in some contexts, and that, when it does, it can have a direct causal effect on moral judging.

The central issue with regard to their respective views of moral reasoning, therefore, does not hinge on how often they argue it occurs, or how much of a direct causal influence they hold moral reasoning can have on moral judging. Rather, the central issue in assessing these models' views of moral reasoning is what they claim the capacity for moral reasoning consists in, that is, what process or set of processes they maintain constitute the capacity for moral reasoning.[1] As the above quotes indicate, Social Intuitionists view the capacity for moral reasoning as an information transformation process that occurs between two people, and Constructive Sentimentalist view the capacity for moral reasoning as a process of categorization. These are the genuinely unique claims of the Social Intuitionist and Constructive Sentimentalist models of moral judgment with respect to moral reasoning.

Moreover, these are the claims that are intended to provide an alternative to the view of moral reasoning given by the deductive model of moral judgment. Recall that according to the deductive model of moral judgment, moral judging involves disinterested and impartial deductive reasoning moral principles. The problem with

[1] Confusion about what the central issue is occurs on both sides of this debate. Social Intuitionists, for example, argue that they are immune from such criticism because, according to their model, 67% of the links of their model (4 out of 6) involve reasoning (Haidt & Bjorklund, 2008b, p. 249). But certainly how many links involve reasoning does not illuminate the question of whether the model gives the right account of what moral reasoning consists in.

the deductive model of moral judgment, according to Social Intuitionists and Constructive Sentimentalists, is that it requires a set of normative capacities that humans do not actually possess. Social Intuitionist argue that the data on moral judging, and moral reasoning in particular, indicates that it is simply not possible for creatures like us, with minds like ours, to engage in disinterested and impartial deduction from moral principles. For example, moral questions, unlike many descriptive questions, are emotionally charged and often have implications for things that people feel very strongly about, which leads people to search for evidence selectively and with a bias to confirm their already arrived at moral judgments; a bias that is well-confirmed in other domains (Koehler, Brenner, & Griffin, 2002; Kunda, 1990). Thus, Social Intuitionists argue that only those who have had some serious training in some very unnatural ways of thinking about morality can set aside these biases to reason disinterestedly and impartially about moral questions. For most ordinary people, the ideal of disinterested and impartial reasoning in the moral domain is simply unattainable. Indeed, many philosophers have already argued as much (notably, for example, Blum, 1980; Williams, 1973a). Constructive Sentimentalists draw a somewhat stronger conclusion. They hold that it is a conceptual truth that reasoning, which is the affect-free manipulation of propositions, cannot be done with moral concepts, because such concepts are partly emotionally constituted.

Moreover, when considering what the empirical literature actually supports, both Social Intuitionists and Constructive Sentimentalists argue that some empirical literature points a very deflationary and skeptical account of the capacity for moral

reasoning. Aside from the large body of evidence that specifically indicates that emotions can play a large causal role in moral judging and that people are often dumbfounded with respect to their moral judgments—that is, they unable to provide further plausible reasons for them (Cushman, et al., 2006; Haidt, 2001; Hauser, 2006; Hauser, et al., 2007), there is more general evidence that people are expert confabulators of reasons for their judgments and actions after the fact (Gazzaniga, 1995; Nisbett & Wilson, 1977). Taken together, this literature points to the need for a different understanding of what moral reasoning consists in, and both Social Intuitionists and Constructive Sentimentalists develop this insight in different ways.

When assessing the Social Intuitionist and Constructive Sentimentalist account of moral reasoning, then, it is important to recognize that neither the Social Intuitionist or Constructive Sentimentalist model involves the "wholesale rejection of reason" in moral judging, as is sometimes claimed (Bloom, 2010). Rather, it is better to understand Social Intuitionists and Constructive Sentimentalists as rejecting a particular conception of what the capacity for moral reasoning consists in on the grounds that it is empirically inadequate—namely, the conception of moral reasoning given by the deductive model of moral judgment, which views the capacity for moral reasoning as disinterested and impartial deduction from first principles—and offering what each sees as an empirically more adequate account of the processes that constitute moral reasoning. Therefore, the central issues here is assessing is how adequate these views of what the capacity for moral reasoning consists in are, which can be assessed by determining how well they cohere with the broader facts of our moral psychology. This is the second condition of adequacy for any model of moral

judgment—that it be consistent with other facts of our moral psychology—and it is here that the Social Intuitionist Model and the Constructive Sentimentalist model raises significant explanatory worries, I argue, because they incur some significant costs to explanatory adequacy when attempting to account for one particularly robust feature of our moral psychology, namely, moral change.

2. Moral Change

One important and common feature of our moral psychology is that people's moral attitudes, commitments, and judgments change over time. People who were racist come to disavow their racist attitudes. Those who were homophobic come to disavow their homophobic attitudes. People get married or have children, which changes their moral commitments in significant ways; parents and spouses commit themselves to the well-being of others in ways that require them to reorder their other moral priorities. Some people become convinced that a particular moral theory is true, leading them to reconsider and change many of their moral attitudes and commitments. Perhaps more common among philosophers are those who become convinced of the falsity of a moral theory that they once considered true, perhaps by finding that an impartial moral theory of justice or utilitarianism rules out the morally significant partiality of important personal relationships (Williams, 1973a). In other, more specific ways, people come to change their moral judgments with respect to some action or person. For example, many people initially judged that the war in Iraq was morally permissible (even morally required), but they now come to judge it as morally impermissible; or there are those who come to judge that eating meat is morally impermissible though they previously judged that it was entirely permissible.

These are just a few examples, but the important point is that it is a manifest phenomenon of our moral psychology that such changes are not only possible, but common. For shorthand, label this phenomenon "moral change".

What is it about minds like ours that makes it possible for people to change their moral attitudes, commitments, and judgments in these ways? Given the complexities of our moral psychology there are many different possible ways moral change can occur. One possible way is direct social pressure. For example, if a person moves to a new moral environment where their former attitudes are simply not condoned, their attitudes, commitments, and judgments can change (perhaps nonconsciously) in order to fit in. A second way moral change can occur is through a moral epiphany—an "aha" moment where a person's moral attitudes, commitments, or judgments change immediately after a sudden realization. W.V.O. Quine is reported to have had a moral epiphany with respect to his anti-Semitic attitudes. His attitudes changed immediately when he suddenly realized that several Jewish men that he met through his philatelist interests, whom he liked, were just members of the set of all Jewish people, and thus that his anti-Semitic attitudes were wrong.[2] A third way that moral change can occur is when a new or revised moral attitude, commitment, or judgment simply "dawns" on a person, slowly and over time. Such people do not experience an "aha" moment, and are completely unaware that their moral attitudes are changing at all. This sort of experience is common in those situations where a person is in close contact with others with whom they hold

[2] This story is told by Chris Cherniak, as reported by his student Elizabeth Picciuto (personal communication).

unfavorable prejudicial attitudes, and through that close contact, the person's racist attitudes eventually change.[3]

A fourth way moral change can occur, and one that is particularly relevant to the current discussion is moral change that occurs as a consequence of distinctively moral reasoning. In moral reasoning people can deliberate with a myriad of considerations that can lead them to evaluate an action or person in a different way, or to weigh their moral commitments differently, or to come to the considered view that some of their own moral attitudes are wrong. Sometimes the considerations a person deliberates with are theory-dependent, such as those required by a moral theory, but more often the considerations that enter into moral reasoning are various intuitions, commitments, principles, and a conception of oneself as a certain kind of person who takes certain considerations seriously, such as honesty, integrity, or thoughtfulness (McDowell, 1995). In moral reasoning people can come to recognize that some of their moral attitudes, commitments, or judgments are incoherent, or conflict with a particular view of themselves, which can lead them to modify, revise, or reject some of their moral attitudes, commitments, and judgments.[4]

A study of moral vegetarians suggests that school-aged children undergo moral change as a result of distinctively moral reasoning (Hussar & Harris, 2010).

---

[3] Mark Twain's Huck Finn is a good example of this sort of dawning. During the course of his river journey with his aunt's escaped slave Jim, Huck catches glimpses Jim's humanity. Through their interactions, Huck begins to see Jim as more than just a slave, or a black man, but as his friend. There is no single moment where Huck has a moral epiphany; rather they become somewhat begrudging friends, over time. Huck does not extend his newfound appreciation of Jim as a human being to all slaves. He does not experience an overall change in attitude toward slaves, but only specifically towards Jim. However, those who work or live in close contact with those can come to have an overall change in their view towards all members of that group.

[4] Moral reasoning does not necessarily lead to such a change, nor does recognizing some incoherence give people guidance with respect to whether and how they should revise a particular moral attitude, commitment, or judgment.

Hussar and Harris investigated children aged 6-10 years-old who were independent moral vegetarians, meaning that they arrived at the moral judgment that eating meat is wrong independently of the moral judgments expressed by their peers or parents. The aim of the study was to determine what led these children to change their moral judgments with respect to meat-eating independently of their peers and parents. Nearly universally, independent moral vegetarians in the study cited the suffering of the animals as the reason for their judgment that eating meat is wrong. More interestingly, though, is the way many of them considered this fact to be morally relevant: because it is not nice to cause suffering. These children saw themselves as nice people, and as such, causing suffering was, to them, inconsistent with how they viewed themselves.[5]

When moral change is the result of moral reasoning, there are varying degrees in which this change results in stable behaviors. For some people, for example, coming to think and act consistently with their moral reasoning toward people of certain racial groups is achieved by force of habit; overriding their implicit negative racial stereotypes by reflexively deploying their reflective attitudes towards such persons. In most situations their overt behavior is relatively stable, but in situations where their reflexive habits are blunted, such as a night of hard drinking or lack of sleep, the person's implicit attitudes come through (Payne, 2005).[6] Other times, however, moral reasoning can lead to a stable and deep change in a person's affective

[5] Of course one should be careful in putting too much theoretical weight on this one study. As research on moral dumbfounding makes clear, people's self-reports of their reasons are not always reliable guides to the underlying processes of moral judging, or, in this case, of moral change, and it is possible that the real cause moral change in this case is something else entirely, such as, perhaps feelings of sympathy.

[6] For example, Mel Gibson is now infamous for the drunken anti-Semitic rant he delivered when pulled over by the police in July 2010.

and motivational dispositions. For example, Fessler et al. found that moral vegetarians are more likely to avoid eating meat compared with health vegetarians or those who avoid meat because they do not like the taste (Fessler, Arguello, Mekdara, & Macias, 2003). Moreover, the moral vegetarians in Fessler et al.'s study were more likely to experience disgust by the thought of eating meat, and that this affective response came about only after they had come to the moral judgment that eating meat is wrong.[7]

In assessing any proposed model of moral judging, one condition of adequacy is that the model of moral judging must be consistent with the facts of moral change. At the very least, a model of moral judging must not render the phenomenon of moral change mysterious or opaque, or rule out the possibility that moral change can occur in the ways outlined above. It would be even better, though, if a model of moral judging shed some light on the underlying psychological processes of moral change as well. It is here, however, that both the Social Intuitionist and Constructive Sentimentalist account of moral reasoning incur significant costs to explanatory adequacy, which raises questions with respect to the overall adequacy of these models of moral judging.

3. The Social Intuitionist Model

It is important to first get clear on what Social Intuitionists claim moral reasoning consists in. Recall that according to the Social Intuitionist Model moral judgments are largely caused and determined by quick emotional responses. Haidt and Bjorklund refer to these quick emotional responses as moral intuitions, and argue

[7] Fessler et al.'s study, however, does not show that emotions are not causally important in bringing about the change in moral judgment in the first place.

that these moral intuitions cause a moral judgment, which includes the belief in the rightness or wrongness of the observed action (Haidt & Bjorklund, 2008a, p. 188). On this model, a moral judgment is largely, though not completely, determined by moral intuitions. Thus, moral judging is largely a matter of a person's emotional responses, which according to Social Intuitionists, have no basis in reasoning. When people are pressed to provide reasons for their moral judgments, they confabulate *post hoc* reasons after the fact. Moreover, according to the Social Intuitionist Model, moral judging is not simply a process of a person's private cognitions, but how a person's intuitions and judgments affect other people as well, and they argue that the *post hoc* reasons a person may offer for her judgment have the potential to influence the moral judgments of others, either through direct social pressure, or because the reasons force other people to view the situation in a different light, causing in them different moral intuitions and therefore leading to new and different moral judgments, and vice versa. The back and forth of *post hoc* reasons, moral intuitions, moral judgments, and social persuasion constitute, according to Social Intuitionists, moral reasoning. This is certainly not a typical account of moral reasoning, but they argue, "If moral reasoning is transforming information to reach a moral judgment, and if this process proceeds in steps such as searching for evidence and then weighing the evidence, then a pair of people discussing a moral issue meets the definition of reasoning" (Haidt & Bjorklund, 2008a, p. 193).

Underlying this social understanding of moral reasoning is a substantive claim about the nature of deliberation in general, which Social Intuitionists seek to explain in evolutionary terms. According to Social Intuitionists, to understand the *natural*

function of reasoning we must determine how it enhances reproductive fitness, which, they argue, it does by allowing us to track the reputations of others and enhance our own reputation in the eyes of others. They do not specify the relationship between reputation tracking and reproductive fitness, but the probable connection they are after is that good reasoning can allow us to reap the benefits of cooperation and coordination by identifying cooperative partners who are unlikely to renege on agreements, as well as make ourselves more attractive cooperative partners to others by providing plausible reasons for our own actions and judgments. Social Intuitionists thus conclude that the natural function of deliberation is not to track the truth, or even to attempt to discover the truth through arguments, but to enhance a person's reproductive fitness through social persuasion. Applying this conclusion to the moral case, moral reasoning naturally functions to enhance a person's reputation in the eyes of others by producing plausible reasons for his or her moral judgments.

Even though Social Intuitionists maintain that moral reasoning is primarily a social phenomenon, they do allow for some forms of intrapersonal moral reasoning. Most of the time, though, intrapersonal moral reasoning is just social reasoning done with oneself. In private conversation a person can consider *post hoc* reasons for a moral judgment, or point to various features of the situation, which might trigger a new intuition leading to a new and different moral judgment. Only in very rare instances can a person reason to a moral judgment through the "sheer force of logic," that is, in a way that does not rely on moral intuitions and emotions (Haidt & Bjorklund, 2008a, p. 193). This sort of intrapersonal reasoning is rare because it requires years of schooling in specialized disciplines such as science or philosophy

that train people to be more reflective, and to practice an "unnatural mode of human thought" (p. 193). Ordinary people who have not been trained in one of these disciplines will only rarely, if ever, be capable of reasoning in this way, and even if they did, argue Social Intuitionists, it would have little influence on that person's actions.

The Social Intuitionist view of intra- and interpersonal moral reasoning, however, incurs some costs to explanatory adequacy when it comes to the phenomenon of moral change, because it cannot provide a straightforward account of how people's moral attitudes, commitments, and judgments change in response to moral reasoning. Certainly Social Intuitionists can account for some instances of moral change by claiming that when people come to a new and different moral judgment with respect to some action or person that it involves a back and forth of *post hoc* reasons that can cause a change in their moral intuitions leading to a new and different moral judgment. But this is not the only way moral change goes—people also change their broader moral commitments, such as coming to the view that a certain moral theory is correct, or that a particular set of considerations is morally relevant in a way they had not previously considered, for example, that people deserve equal respect regardless of race—which leads to different moral judgments in particular cases. Moreover, people sometimes reason with the aim of subjecting their own moral attitudes and judgments to reflective scrutiny to assess whether their moral attitudes, commitments, and judgments cohere in the right sort of way with each other, with other non-moral facts, and with a particular view of themselves. Most people probably do not undertake a project of global coherence, but rather reflect on

much more limited local inconsistencies, such as their moral attitudes towards meat-eating and the suffering of animals, or their prejudicial attitudes towards members of different racial groups and their view that all humans deserve equal respect. This is one way racists come to disavow their racist attitudes, and six-year-olds become moral vegetarians.

The problem with the Social Intuitionist account of moral reasoning when it comes to accounting for such instances of moral change is that, according to Social Intuitionists, people cannot reason with the aim of achieving coherence among their various moral attitudes, commitments, and judgments. But this claim implies that people can be subject to a deep sort of moral incoherence, because their moral judgments are caused by a set of nonrational and potentially inconsistent moral intuitions, and they lack any reflective capacity for bringing these judgments into greater coherence with each other or their other moral attitudes, commitments, and non-moral beliefs.[8] On this view, it is possible for a person to be in a position where she simply cannot help but judge that eating meat is permissible, even if she finds such a judgment to be deeply incoherent with her other moral commitments or how she views herself. Such a situation would be a sort of moral incoherence in just the same way that being unable to revise some other attitude that is deeply incoherent with one's other beliefs is a sort of epistemic incoherence. But people are not morally incoherent in this way, at least not typically (though many people are sometimes

[8] It might be possible for Social Intuitionists to give some non-reflective account of how people can achieve some greater coherence among their moral judgments through, perhaps, imagining actual and counterfactual examples to produce new intuitions and judgments, which perhaps overturn other intuitions and judgments. But the difficulty for this sort of account is that for Social Intuitionists, one's other moral attitudes and commitments cannot be brought to bear in evaluating one's moral judgments. I shall return to this point in my discussion of Constructive Sentimentalism, and will develop this problem more fully.

deeply inconsistent in their moral judgments), and so it is possible for people to modify, revise, or reject their moral attitudes, commitments, and judgments in light of their moral reasoning and a desire for some degree of coherence among their moral attitudes, commitments, judgments, and other non-moral facts.

Moreover, even if the Social Intuitionist Model were to explain how people could avoid moral incoherence through reasoning, Social Intuitionists would have to show that considerations of coherence and considerations with respect to some non-moral facts are connected to the emotions in the right sort of way. On the Social Intuitionist Model, a consideration can gain a grip in moral thinking only if it is connected to the emotions in the right sort of way that can potentially lead to a new and different moral intuition. But it is extremely difficult to see how a desire for coherence among one's moral judgments, or a specific dimension of consistency, is connected to the emotions in the right sort of way to lead to a new and different moral intuition. For example, it is extremely difficult to see how coming to see that race is a morally irrelevant difference can lead to a new and different moral intuition. And yet such considerations do lead people to modify, revise, or reject some of their attitudes, which can lead to new and different moral judgments and actions.

The problems I have raised for the Social Intuitionist Model are not a knockdown argument—it does not show that the Social Intuitionist Model is wrong—but it does show what Social Intuitionists claim moral reasoning consists in does not cohere with broader facts of our moral psychology, because it is not consistent with some aspects of moral change. There are thus some elements of the Social Intuitionists Model that fail the second condition of adequacy that any model of moral

judging must satisfy. This does not yet provide sufficient reason to dismiss the Social Intuitionist Model, but it does provide sufficient motivation to look to alternative explanatory models of moral judging and alternative accounts of moral reasoning, such as the Constructive Sentimentalist model.

4. The Constructive Sentimentalist Model

Constructive Sentimentalists are even more skeptical of the role of moral reasoning in moral judging than Social Intuitionists, and their very narrow view of the role of reasoning in moral judging raises a number of explanatory worries when it comes to explaining moral change as well. Constructive Sentimentalists maintain that moral reasoning only ever amounts to either *post hoc* justification of already arrived at moral judgments, or deliberation to determine whether some action or person falls under a concept for which a person has a moral rule. According to Constructive Sentimentalists, it is a conceptual truth that moral reasoning is limited in this way because moral concepts, such as WRONG, are "a detector for the property of wrongness that comprises a sentiment that disposes its possessor to experience emotions in the disapprobation range" (Prinz, 2007, p. 94). From this analysis of moral concepts, Constructive Sentimentalists argue that moral attitudes, commitments, and judgments are constituted by sentiments and emotions, and because moral reasoning is the affect free manipulation of propositional attitudes (p. 99), it follows that moral reasoning is severely limited in its possible causal roles in moral judging. On this view, moral reasoning does not and cannot involve deliberation *with* one's own moral attitudes, commitments, and judgments, because those attitudes, commitments, and judgments are constituted by one's sentiments and

emotions. The most people can do in moral reasoning is deliberate about non-moral facts to determine whether some object falls under a concept, or provide post hoc reasons for their emotionally constituted moral judgments.

If moral reasoning is limited in the ways that Constructive Sentimentalists argue, then how can they account for moral change that results from moral reasoning? For Constructive Sentimentalists, moral change can result from moral reasoning when, through moral reasoning, a person determines that some action or person is better thought of as falling under a different concept towards which that person has a moral sentiment. A person can deliberate with a range of non-moral considerations, draw analogies with other cases, or highlight some features and diminish others, which can favor categorizing an action or person as falling under a different concept. This new categorization of an action or person can then lead to a different moral judgment if that person has a different sentiment towards that concept, because the person will be disposed to feel differently about it, that is, disposed to judge it differently.

Prinz describes how moral reasoning can lead to moral change by drawing an analogy with how our judgments of other people can change through reasoning (Prinz, 2007, pp. 122-125). Suppose Smith is trying to convince Jones that another person is likeable, though Jones initially judged that this person was boorish and arrogant. How would Smith go about convincing Jones to change his judgment? Most likely, Smith would diminish this person's bad qualities, and accentuate the positive ones. Smith would provide alternative explanations for this person's seemingly boorish and arrogant behavior, for example, that new people make him nervous.

Smith would point out that this person has a good sense of humor, is loyal, and kind, and similar other things. So long as Jones has positive sentiments toward the traits that Smith points out, Jones's initial judgment may begin to change. Jones will not automatically feel differently, argues Prinz, but Jones will be disposed to feel differently, that is, be disposed to judge the likeability this person differently.

For Constructive Sentimentalists, moral change as a result of moral reasoning is brought about in just the same way. In moral reasoning a person deliberates with considerations for categorizing an action or person under one concept or set of concepts instead of another, and perhaps corrects mistakes with respect to some set of nonmoral facts. If the considerations lead a person to categorize an action or person differently, under a concept towards which that person has a different sentiment, then the person will be disposed to feel differently, and therefore disposed to judge differently. On this view, moral reasoning does not involve thinking with or about one's own moral attitudes, commitments, or judgments, but rather it involves reasoning with the aim of discovering whether some object *x*, is an instance of some *F* or some *G*, such that *F* and *G* are concepts towards which one has a sentiment. If, in reasoning a person concludes that *x* falls under the concept *F,* though he previously held that it fell under the concept *G*, and the person has a different sentiment towards *F* than he does to *G*, then that person's moral judgment will be disposed to change.

An example from Prinz might make the scope and limits of moral reasoning on this view clearer. Prinz writes:

> Suppose that I say that prayer in school is wrong. I might add a reason: it is wrong because it discriminates against atheists and members of

> minority religions. You might reply that it does not discriminate. Or you might reply that prohibitions against school prayer discriminate against members of the majority religion, and hence discriminate against more than if the prohibitions were lifted. This is a rational debate. We could settle on a prior definition of discrimination, and provide evidence for our respective views. If I persuade you of my view, your emotional attitude toward prayer in school would not change instantly, but it would be disposed to change. But now suppose that you are not persuaded. Suppose instead that you say discrimination is not morally wrong. The best I can do is stare at you incredulously. If you think discrimination is not wrong, then we are constituted differently (Prinz, 2007 pg. 124).

What is important to stress here is that according to Constructive Sentimentalists, it is not possible for people to engage in any distinctively moral forms of reasoning at all. That is, in moral reasoning people do not consider whether some action is right or wrong, or whether some person is good or bad, because this would involve reasoning *with* one's moral attitudes, commitments, and judgments, which are emotionally constituted on this view. Reasoning can only involve the affect-free manipulation of propositions, so when people undertake to answer a moral question they can only consider whether some *x* is some *G*, or some *F*, or has more *G*-type properties than *F*-type properties, which is supposed to be a purely *descriptive*

question. The question cannot have any moral content, because moral content is emotionally constituted.

It is because reasoning is limited in this way to purely descriptive content on this view that the Constructive Sentimentalist model of moral judging and moral reasoning raises some explanatory worries when it comes to moral change. First, not all aspects of moral reasoning that leads to moral change can be adequately explained as purely descriptive reasoning. Many morally relevant concepts, that is, those concepts that are likely to figure in episodes of moral reasoning, such as murder or stealing, are not purely descriptive, but rather, contain some distinctively moral content. Murder is not just killing; it is wrongful killing. Stealing is not just taking; it is wrongful taking. To determine whether some action is rightfully categorized as an instance of murder, for example, it is not possible simply to ask whether it satisfies certain descriptive features of killing, because determining whether a killing is a murder requires answering the further distinctively moral question of whether the killing is wrongful or justified. That is, a person must have some notion, however vague and inchoate, of the sorts of conditions that serve to justify a killing, and the sorts of conditions that do not, in order to categorize any killing as a murder. But these conditions are not purely descriptive; they are moral. Thus, it is not always possible to determine whether some *x* is better categorized as some *G* or some *F* without reasoning in distinctively moral ways with moral content.

This is clear even in the example that Prinz provides. Prinz argues that in moral reasoning a person can consider reasons for determining whether school prayer is wrong by considering whether it is a case of discrimination, and if it is a case of

discrimination, then it is wrong. However, it is not the case that discrimination *tout court* is wrong. There are many permissible instances of discrimination, such as discriminating amongst candidates for admission based on their grade-point-averages or standardized test scores. If school prayer is wrong because it is a form of discrimination, it is only if it is a wrongful sort of discrimination, that is, if it discriminates among people according to morally irrelevant or suspect grounds. But one cannot determine whether some action or policy is permissibly or impermissibly discriminatory without reasoning with distinctively moral content with respect to whether the grounds of discrimination are morally irrelevant or suspect.

Moreover, even if it were the case that some morally relevant categorizations could be settled on purely descriptive grounds, it is not the case that such categorizations by themselves always settle moral questions. Take, for example, the question of whether it was wrong to waterboard Khalid Sheik Mohammad 183 times in one month.[9] On the Constructive Sentimentalists view of moral reasoning, a debate between those who hold different moral views on this question might focus on whether waterboarding is better thought of as an enhanced interrogation technique, and therefore permissible, or whether it is better thought of as torture, and therefore impermissible. They may settle on a prior definition of enhanced interrogation techniques and torture, and argue about how various descriptive aspects of this prolonged action fall under one or another of those categorizations. Now, suppose that these interlocutors decide that, on purely descriptive grounds, waterboarding Khalid Sheik Mohammad 183 times in one month counts as torture. Assuming that

[9] This information came to light in August 2009 with the public release of a heavily redacted 2005 Department of Justice Memo (May 30, 2005).

these interlocutors share a sentiment against torture, does this categorization settle the moral question? No, because there are some who hold that the waterboarding of Khalid Sheik Mohammad was torture, but it was *justified* torture (Krauthammer, 2009). So again, the moral question of whether and when waterboarding may be permissible is not settled simply by categorizing it as an enhanced interrogation technique or torture. Even if it is torture, some people argue that it is nonetheless permissible in certain situations, such as when a terrorist knows the location of a ticking time bomb (Dershowitz, 2002). The point is that it is not the case the descriptive categorizations settle moral questions, nor is it the case that seemingly descriptive categorizations are always settled free of moral considerations.

Thus, the Constructive Sentimentalist account of moral reasoning, which limits moral reasoning to purely descriptive reasoning, renders some aspects of moral change quite opaque because they cannot be adequately explained by claiming that people come to change their moral judgments simply by changing their purely descriptive categorizations of an object of appraisal. Sometimes moral change involves determining that an object of appraisal is still an *x*, but a justified or unjustified instance of *x*; a determination which requires reasoning *with* moral concepts, such as good, bad, permissible, and impermissible. Moreover, even if an instance of moral change does involve a recategorization of an object of appraisal under a different concept, the concepts involved often contain moral content as well. So, even in some instances where moral change is brought about by recategorizing an object of appraisal, that reasoning is not always purely descriptive. Constructive Sentimentalists cannot account for these two aspects of reasoning that leads to moral

change, because they are committed to the view that reasoning is always purely descriptive. This is a significant explanatory cost to the model.

A second general explanatory worry for the Constructive Sentimentalist account of moral reasoning and moral change is that, on this view, the underlying moral attitudes and commitments that give rise to a person's moral judgments can never themselves be the object of moral reasoning, and thus a person can never seek to change their underlying moral attitudes or commitments as a result of moral reasoning. On this view, a person's moral judgments can change as a result of reasoning, but their underlying moral attitudes and commitments cannot. On this view, then, it is not possible for a person to consider whether their attitude toward discrimination, for example, is appropriate or coheres in the right way with their other moral attitudes, commitments, judgments, and non-moral beliefs and to seek to change or modify that attitude it if it does not. This is because, according to Constructive Sentimentalists, a person's basic moral attitudes and commitments are brute psychological facts of that person that lie beyond the scope of reasoning. Prinz writes that, "basic values [such as moral attitudes] are implemented in our psychology in a way that puts them outside certain practices of justification," which includes determining whether they cohere with other of one's moral attitudes, commitments, and judgments (pg. 32).[10] This problem applies, *mutatis mutandis*, to the Social Intuitionist Model as well.

[10] One natural way of thinking Constructive Sentimentalists could account for such reasoning is to claim that a person can consider the *appropriateness* of an emotional response to an object of evaluation in moral reasoning. Indeed, many neo-sentimentalists analyses of moral concepts build in just this sort of meta-cognition, in part, to account for the fact that people often do think about whether their moral judgments cohere with their other moral attitudes and commitments and beliefs about the world (D'Arms & Jacobson, 2000b; McDowell, 1988a; Wiggins, 1987). On these analyses, in moral reasoning a person can consider whether a particular emotional response is coherent with their other

The problem for Constructive Sentimentalists is that it is not uncommon or unusual for people to conclude that some of their own moral attitudes and commitments are mistaken or inappropriate in virtue of the fact that those attitudes and commitments do not cohere in the right way with other of their moral attitudes, commitments, judgments, and non-moral beliefs. Indeed, racists come to disavow their racist attitudes after concluding that their own racist attitudes are inappropriate; or homophobes come to disavow their homophobic attitudes after concluding that their own homophobic attitudes are inappropriate. Moreover, not only do people come to disavow some of their own moral attitudes and commitments, they often also seek to change or modify their behavior as a result of such reasoning, such as refusing to participate in forms of joking they now consider inappropriate. This is a perfectly common form of moral change, and the problem for Constructive Sentimentalist accounts of moral reasoning and moral change is that it renders mysterious how it is people do in fact reason in this way with their moral attitudes and commitments, can come to view some of them as inappropriate, and change their behaviors as a result.

There are at least two significant explanatory worries that the Constructive Sentimentalist model of moral judgment and moral reasoning raises with respect to moral change. And again, the point of raising these worries is not to provide a knockdown argument against the Constructive Sentimentalists model of moral

---

emotional responses, or non-moral facts. Moral reasoning is no different from the sort of reasoning an arachnophobe undertakes to determine that his terror towards spiders is unwarranted because it is inconsistent with the actual facts of spiders. However, Constructive Sentimentalists flatly reject this view of moral reasoning (pp. 111-115). Prinz argues that sentimentalists who build in this sort of meta-cognition are committed to the view that moral judgments are judgments about the appropriateness of an emotional response, rather than judgments constituted by emotional responses. And this meta-cognitive analysis of moral judgments, argues Prinz, is inconsistent with the Constructive Sentimentalist analysis of moral concepts, which are constituted by emotions, not norms with respect to the appropriateness of an emotion.

judging. Rather, the aim is to show that how Constructive Sentimentalists view what moral reasoning consists in is not consistent with one particularly common aspect of our moral psychology, and indeed, renders it utterly mysterious or opaque. There are thus some elements of the Constructive Sentimentalist model that fail the second condition of adequacy that any model of moral judging must satisfy. This provides sufficient motivation to look to alternative explanatory models of moral judging and alternative accounts of what moral reasoning consists in. In what follows, I shall offer my own account of what moral reasoning consists in, which I argue is consistent with our actual normative capacities and can provide an account for how distinctively moral reasoning can lead to moral change. In later chapters I shall incorporate this view of moral reasoning into an overall framework for moral judging.

5. Moral Reasoning, Naturalized

The Social Intuitionists' and Constructive Sentimentalists' views of what the capacity for moral reasoning consists in have some problems accounting for a central fact of human moral psychology, which raises questions with respect to the overall adequacy of their respective models of moral judging. However, it is not uncommon for any relatively new explanatory model to have lingering questions or problems that are offset by other theoretical considerations, such as how well it unifies diverse phenomena, or how simple and elegant it is. It is not sufficient, therefore, simply to show that the Social Intuitionist and Constructive Sentimentalist models of moral judging have some theoretical and explanatory problems; one needs to offer an alternative explanatory model of moral judging that fares better along these dimensions. That is a much larger project that shall take the rest of the dissertation to

develop, but what I want to do here is to offer an alternative naturalized account of the capacity for moral reasoning that does not have the same explanatory problems with respect to moral change that beset both the Social Intuitionist and Constructive Sentimentalist models of moral judging.

To begin, Social Intuitionists and Constructive Sentimentalists are correct in emphasizing that moral reasoning is not a transcendental and disembodied capacity that involves wholly abstract deduction from first principles as the deductive model of moral judgment maintains. Indeed, there is little reason to think that ethics can permit of deductive proof, as the deductive model claims, especially when so few other human cognitive endeavors do. For example, scientists do not produce deductive proofs of natural laws or theories; rather, they build a case for them using the best available evidence, methods, and related theories. If scientific reasoning does not proceed deductively, then why should there be any expectation that moral reasoning can proceed deductively, especially since questions of how we are to live and what we ought to do are sometimes more complicated than scientific ones? As Aristotle put it, we should only expect a level of precision consistent with the enterprise,[11] and the level of precision possible for ethics is not that of a geometric proof, because the starting points for moral reasoning are rarely, if ever, abstract first principles of morality, but rather it is from within the point of view of a particular moral agent, with a certain history, socialization, beliefs, and experiences (Flanagan, 1991, p. 53;

[11] Aristotle writes in the *Nicomachean Ethics*:

> Our discussion will be adequate if it has as much clearness as the subject-matter admits of, for precision is not to be sought for alike in all discussions, any more than in all the products of the crafts…for it is the mark of an educated man to look for precision in each class of things just so far as the nature of the subject admits; it is evidently equally foolish to accept probable reasoning from a mathematician and to demand from a rhetorician scientific proofs (Aristotle, 1999, pp. Book I, Section 3).

Nagel, 1986). What is needed, then, is a naturalized account of what the capacity for moral reasoning consists in, just as Social Intuitionists and Constructive Sentimentalists maintain, but in a way that avoids the explanatory problems that come with adopting either of those views.

A better way to naturalize moral reasoning is in just the way that has been implicit throughout this chapter. On this view, moral reasoning takes place within the particular point of view an agent with particular experiences, intuitions, beliefs, and some (possibly inchoate) moral-theoretic considerations, and consists in subjecting one's moral attitudes, commitments, and judgments to reflective scrutiny to achieve broad coherence among them and with one's experiences, intuitions, and non-moral beliefs; and then modifying, revising, or rejecting some of one's moral attitudes, commitments, and judgments in order to achieve (greater) coherence. In moral reasoning one can consider arguments offered by others, but they will be evaluated against the backdrop of that particular person's experiences, intuitions, and non-moral beliefs.

This naturalized account of what moral reasoning consists in is in keeping with the method of reflective equilibrium, in particular, of wide reflective equilibrium (Daniels, 1979; Rawls, 1999). In the process of wide reflective equilibrium, a person seeks coherence among their various moral judgments, explicit moral principles, and other beliefs, such as the sort of person they believe themselves to be, by revising or modifying any of these particular elements to achieve a broad coherence among them all. The end-point of this process is a reflective equilibrium among moral judgments, explicit moral principles, and other beliefs. There are important questions about

whether the process of reflective equilibrium justifies a person's moral judgments or explicit moral principles. We shall return to that question in the next chapter, but the point here is that the process of wide reflective equilibrium provides a non-deductive picture of what moral reasoning consists in consistent with the normative capacities we actually possess. On this view of what moral reasoning consists in, people do not reason deductively from first principles, but instead reason to get their various moral attitudes, commitments, judgments, and non-moral beliefs to hang together in the right sort of way.

A few caveats are necessary, though, in order to fill out this account of moral reasoning. First, it may not always be possible for a person to revise or modify a moral attitude, commitment, or judgment directly. Many of our moral attitudes and commitments are nonconscious, and are not directly corrigible by conscious reflection. In such cases, a person may have a dual-attitude, where their initial moral judgments in a range of cases conflict with their settled moral view. This often happens in the case of racist attitudes. Many racial stereotypes that give rise to various judgments are implicit attitudes that are not directly corrigible by conscious reflection. A person does not stop making racist judgments about people simply by recognizing that their racist judgments conflict with their commitment that all people deserve equal respect (Quine, perhaps, excepted). They may have to work hard to override these initial judgments consciously using higher-level executive processes, though this override may become habitual enough that it often no longer takes conscious intervention to override these initial judgments. However, when higher-level executive processes become attenuated, such as after a hard night of drinking, a

person's initial racist judgments may not be overridden, and the person may judge and act consistently with them. Actually changing the underlying implicit racial attitudes requires more indirect means, such as close contact with members of that group, or reading stories where members of that group are the heroes.[12]

Second, moral reasoning so conceived will likely be piecemeal, undertaken only when time permits and local inconsistencies are noticed. On this latter point, social discourse is actually quite important. Other people are often quite better at noticing our inconsistencies than we are, and in conversation others can help us see our own inconsistencies much better than we can on our own. Moreover, in social discourse, including reading articles and books, others can offer arguments that challenge our moral attitudes, commitment, and judgments. People will assess the strength of those arguments from within their own point of view, including their other moral commitments, attitudes, judgments, and non-moral beliefs. Deductions from first principles of morality may not convince many people to change their minds if the conclusion conflicts with their other commitments, attitudes, and judgments. This is to be expected, though, because people undertake moral reasoning with the aim of assessing how well all these elements of their moral psychology hang together, and if a supposed first principle of morality conflicts with these other elements, it is the principle that is likely to be rejected, not their other moral commitments, attitudes, and judgments.

Third, it is important to distinguish among three things: giving an account of what a capacity consists in, people's dispositions to engage in that capacity, and the

[12] Paul Bloom reports that a graduate student of his, upon discovering that he was heavily implicitly biased to favor members of his own race, undertook just such steps until measures of implicit bias no longer indicated a strong preference for members of his own race (Bloom, 2007).

normative standards for determining whether someone has engaged in that capacity well. Giving an account of what a capacity consists in, that is, what processes and procedures make it up, does not entail any claims with respect to how often people engage in that capacity, nor how well they do so when they do. People might have a capacity that they engage in only rarely, and when they do engage in it, they do so poorly. Some individuals may only rarely, if ever, engage in moral reasoning. There are people who rarely subject their own attitudes and judgments to reflective scrutiny across a range of domains (Baron, 1985; Stanovich, 2009; Stanovich & West, 1988), including their moral attitudes and judgments. But it is important to distinguish between a capacity and a disposition to use that capacity. Lacking a disposition to engage in moral reasoning, even if widespread, is different than a particular capacity being rare. Some people engage in such reflective thought often, others engage in reflective thought only very rarely, and most people occupy the wide middle ground. But even those who engage in reflective thought more often require some exogenous factors to be able to do it, such as time and the absence of other cognitive tasks. It is therefore, not very much of a surprise that in an experimental setting people rarely engage in any sort of moral reasoning; which perhaps explains why psychologists and neuroscientists find it easy to ignore in developing models of moral judging.

There is, therefore, no straightforward answers to the questions of how often people engage in moral reasoning, or how often it has a causal role in moral judging. This will be different for different people, depending, in part on how often they engage in moral reasoning and what training they may have received, which again, is why the central issue here is not how often moral reasoning has a causal role in moral

judgment, which varies from person to person, but what the capacity moral reasoning consists in, which involves a stable cluster of abilities across ordinary people.

The account of moral reasoning outlined here is a naturalized account of moral reasoning, and can easily account for the facts of moral change. Moral change occurs when people reflect on their moral attitudes, commitments, and judgments to determine how well they cohere with each other, and then reject, modify, or revise some of them to achieve greater coherence. Moreover, this naturalized account of moral reasoning has no difficulty in explaining the Social Intuitionist's claim that people often "feel" their way to a moral conclusion—meaning that there is a back and forth between that person's intuitions, judgments, and explicit moral principles. Certainly the account of moral reasoning I have provided is consistent with the observation that moral reasoning involves a back and forth among many different elements in a person's moral psychology, their various moral attitudes, commitments, and judgments, but this hardly implies that moral reasoning is entirely *post hoc*. Certainly there are occasions where people will simply attempt to rationalize their moral judgments, but that is not the whole story. And because Social Intuitionists limit moral reasoning in this way, their model of moral judging is inconsistent with wider facts of our moral psychology. The account of moral reasoning I provide does not face the same difficulty, and indeed, fits with the broader and complex facts of our moral psychology quite well.

My account of moral reasoning is preferable to those of Social Intuitionists and Constructive Sentimentalists because it can easily account for the facts of moral change, and indeed, helps to explain them. It is preferable to the Constructive

Sentimentalist account of moral reasoning in that my account of moral reasoning can capture the fact that great deal of moral reasoning is distinctively moral in character—that is, it involves reasoning with moral concepts. It is also preferable to the Social Intuitionist account of moral reasoning because Social Intuitionists fail to distinguish between moral reasoning as a capacity and people's dispositions to engage in moral reasoning. There is all sorts of evidence that people, in ordinary circumstances, reason with so-called "myside bias," (Stanovich, 2009) and this is just as true in the moral case. However, from this it does not follow that people generally lack a capacity to engage in moral reasoning reflectively, which is what Social Intuitionists claim.

6. Conclusion

The Social Intuitionist's and Constructive Sentimentalist's account of what moral reasoning consists in face serious difficulties in explaining central facts of our moral psychology, in particular, the facts of moral change. However, both Social Intuitionists and Constructive Sentimentalists are right to insist that a naturalized account of moral reasoning, that is consistent with the actual reasoning capacities of creatures like us, is crucial for developing an empirically adequate model of moral judging. I have argued that an alternative account of naturalized moral reasoning provides a better explanation of the facts of moral change, where moral reasoning consists in attempting to achieve a broad coherence among one's moral attitudes, commitments, and judgments. This account captures the central insights of the Social Intuitionist and Constructive Sentimentalist models, but avoids their skeptical and deflationary conclusions with respect to the role of reasoning in moral judging.

Chapter 4: Can Moral Judgments Be Justified?

New wave sentimentalists argue that there is a further implication of their respective views of the role of moral reasoning in moral judging, namely, that moral judgments are not, and cannot be, justified by reasons. If moral judgments are caused or constituted by the emotions, and these emotional reactions have no connection to reasons or reasoning, then, they conclude, moral judgments are not rationally assessable judgments. They are not the sorts of things that can be appropriate or inappropriate, and they are not the sorts of things that can be rationally supported by reasons. As Prinz puts it, moral judgments are caused in ways "that puts them outside certain practices of justification" (Prinz, 2007, p. 32), and that attempts to justify them are a "fool's errand" (pg. 125). Similarly, Haidt and Bjorklund write that moral judgments are not "justifiable by reasons…" (Haidt & Bjorklund, 2008b, p. 250), and that "reasoning is not a firm enough foundation upon which to ground a theory—*normative* or descriptive—of human morality (Haidt & Bjorklund, 2008a, p. 216, emphasis added). Social Intuitionists and Constructive Sentimentalists argue that these skeptical conclusions follow *directly* from their models of moral judging.

The aim of this chapter is to show that this skepticism with respect to the justifiability of moral judgments does not follow *directly* from either Social Intuitionists' or Constructive Sentimentalists' respective models of moral judging. Importantly, investigating their epistemic claims will help to further illuminate the structure of moral reasoning and its role in our broader moral psychology, and will show that the naturalized account of moral reasoning I developed in the previous

chapter provides a better account of what moral reasoning consists in than either of those offered by Social Intuitionists and Constructive Sentimentalists. This provides further reason for concluding that new wave sentimentalists have given the wrong account of what moral reasoning consists in, and provides further theoretical pressure to reject the claim of new wave sentimentalists that recognizing an important causal role for the emotions in moral judging requires minimizing the role of moral reasoning.

1. The Regress Argument

Both Social Intuitionists and Constructive Sentimentalists argue that their respective accounts of moral judging, which severely limit the causal roles of moral reasoning, have *direct* normative implications. Specifically, they argue that their respective models of moral judging entail that moral judgments cannot, in any meaningful sense, be justified by reasons.[1] As Prinz puts it, the Constructive Sentimentalist model of moral judging reveals that "basic values are implemented in our psychology in a way that puts them outside certain practices of justification" (Prinz, 2007, p. 32). Similarly, Haidt and Bjorklund write that moral judgments are not "justifiable by reasons…" (Haidt & Bjorklund, 2008b, p. 250), and that "reasoning is not a firm enough foundation upon which to ground a theory—*normative* or descriptive—of human morality (Haidt & Bjorklund, 2008a, p. 216, emphasis added). Thus, both Social Intuitionists and Constructive Sentimentalists

[1] What counts as a reason is a hotly contested topic, and there are many substantive accounts with respect to what reasons are, how they are constituted, and how people can reason with them. I am not concerned with these debates here, and I do not use the term reason in any substantive sense. By reason I mean simply any consideration that can enter into reasoning, and by moral reason I mean simply any consideration that can enter into moral thinking.

claim that their respective models of moral judging have direct consequences with respect to the epistemic project in ethics; namely, their respective models of moral judging show that it is a "fool's errand" because moral judgments cannot be justified by reasons.

This is a very skeptical result, and a surprising one at that. A model of moral judging is primarily meant to provide an explanation for a capacity, namely, the capacity of ordinary mature people to make moral judgments, while the epistemic project in ethics, on the other hand, is primarily meant to provide moral guidance, that is, to recommend certain attitudes and actions because they are morally justifiable (Held, 1996). There is an important difference in the subject matter between explanations and recommendations, and, at least initially, it is unclear how psychological claims *could* have direct epistemic implications given that psychological and epistemic claims are about different sorts of things. As Held writes:

> If moral psychology is the psychology of making moral judgments and developing moral attitudes, it seeks causal explanations of how this is done. This leaves unaddressed the normative questions of whether the positions arrived at are morally justifiable" (Held, 1996, p. 70).

This is a contemporary restatement of Hume's famous observation that it is not possible to derive an "ought" from an "is."[2] Both Social Intuitionists and Constructive Sentimentalists, however, directly challenge the view that psychological claims by themselves have no direct epistemic implications. They both argue that, in

[2] *A Treatise of Human Nature* (3.1.1.27).

general, it is possible to derive an "ought" from an "is," at least in the sense of showing that the applicability of some "ought" claims can be ruled out given certain psychological "is" claims (Haidt & Bjorklund, 2008a, pp. 213-217; Prinz, 2007, pp. 1-10). More specifically, they argue that their respective models of moral judging rule out the possibility that moral judgments can be justified. It is this specific argument that shall be my focus.

Social Intuitionists and Constructive Sentimentalists offer structurally similar arguments for the conclusion that moral judgments cannot be justified. Prinz gives a much more straightforward presentation of the argument, and so I will outline his first, and then show how the same argument is implicit in a number of claims Social Intuitionists' make with respect to moral justification.

Constructive Sentimentalists argue that moral judgments cannot be justified, because ultimately, moral judgments are not *based* in reasons. Why is that? Constructive Sentimentalists claim that a reason is an answer to a "why" question (Prinz, 2007, pp. 31-32, pg. 124). For example, if Jones gets a drink of water and we ask him why, his answer to that question is his reason for getting a drink of water. We can keep asking him why questions, but at some point he will no longer be able to provide answers for them, which means that he has no further reasons for getting a drink of water. For Jones it might be that he felt thirsty, or maybe he is trying to avoid future discomfort (if, for example, he needs to stay hydrated for an upcoming run, because knowing if he does not he will develop cramps). Either way, we will have

reached something basic in Jones's psychology for which no further reasons can be given.[3]

Similarly, argues Prinz, if someone judges that some action is morally wrong, then the answers the person gives in response to a series of why questions for that moral judgment are that person's reasons for it. A person will be able to give some reasons for their moral judgments—reasons of categorizing that action as falling under a concept towards which that person has a sentiment, for example—but the reasons that a person can give for a moral judgment do not terminate in well worked out arguments or first principles of morality, or anything of the sort (something demonstrated rather well with research on moral dumbfounding). Rather, the reasons people can offer for their moral judgments usually terminate in some basic moral claim, such as "murder is wrong," or "incest is wrong," for which no further reasons can be given. Prinz writes:

> It's not the case that I value human life because of some well worked out rational argument, and I don't feel any obligation to generate such an argument…If I encounter someone who baldly states that human life has no value, I assume that the person is depraved, not dumb. I respond, not with reason, but with the fist (pg. 125).

Ignoring the rhetorical flourish, Prinz's point is clear enough: the reasons people offer for their moral judgments do not terminate in well worked out arguments, but rather, they terminate in substantive, basic moral claims for which no

[3] Something is psychologically basic, in this sense, if it is where the head-internal explanation of some judgment terminates.

further reasons can be given. Prinz calls such basic moral claims "grounding norms" (pg. 125), and importantly, these grounding norms are themselves beyond the scope of rational criticism because they themselves have no rational basis. There are literally no reasons for them; they are simply brute facts of our psychology. As Prinz writes, "Basic values [grounding norms] provide reasons, but they are not based in reasons" (Prinz, 2007, p. 32). Call this the Regress Argument.

What the Regress Argument is meant to show is that, at bottom, moral judgments are caused by something that is not itself based in reasons, and therefore that moral judgments are not and cannot be connected to reasons in the right way, and therefore that moral judgments are *incapable* of justification. Because the ultimate causes of moral judgments cannot be justified, that is, based in reasons via well worked out arguments, then those judgments that follow from such grounding norms are not themselves, in any meaningful sense, justified by reasons. This is argument is not meant to imply that some moral judgments might not be *better* than others, in a more limited sense. For example, on the Constructive Sentimentalist view it is possible that one might have better reasons for categorizing school prayer as an instance of discrimination rather than as, say, an instance of free speech, and assuming that a person has a sentiment against discrimination, there is a sense in which the moral judgment that school prayer is wrong is better than the judgment that it is permissible. But that is the extent to which reasons can provide rational support for a moral judgment: they do not go beyond recommending or requiring categorizing some object in a particular way.

The Regress Argument can be schematized this way:

1. The reasons for a moral judgment terminate in a grounding norm.
2. Grounding norms are not based in reasons.
3. Therefore, moral judgments are not ultimately based in reasons.
4. Therefore, moral judgments ultimately cannot be justified by reasons.

If correct, this argument derives an "ought" from an "is," in the sense that it rules out the applicability of certain "ought" claims on the basis of a set of psychological "is" claims. If correct, the psychological does indeed have direct implications for moral epistemology, in that it shows that moral judgments cannot be justified given how moral judgments are caused.

Although Social Intuitionists are less clear about it, they too can be interpreted as offering a version of the Regress Argument. The primary motivation for the Social Intuitionist Model is research with respect to moral dumbfounding. Haidt and colleagues observed that people are able to make quick moral judgments, but often struggle to provide reasons for them beyond an accepted cultural rule. When pressed to provide reasons beyond such cultural rules, most people are unable to articulate any further plausible reason (Haidt, 2001; Haidt, et al., 2000; Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993). In that sense, they are morally dumbfounded: they cannot provide any further reason for their moral judgment. Importantly for Social Intuitionists, accepted cultural rules are basic in a person's psychology in that no further reasons can be offered for them. Moreover, people do not accept cultural rules on the basis of some well worked out argument; they accept these cultural rules because all ordinary people have a set of innate emotional dispositions that make

them sensitive to the moral rules in their culture, and, which more often than not, cause people to accept the rules of their culture.[4] Moreover, these cultural rules are not derived from well worked out arguments either—they are developed over time as a means to compel in-group stability and conformity, and to allow members to recognize and mark outsiders easily.

Because accepted cultural rules are psychologically basic, people cannot engage in a rational debate with respect to their accepted cultural rules. This can be seen most easily when thinking of members of different cultures, according to Social Intuitionists. For example, Haidt and Bjorklund argue that we in the West cannot engage in rational dispute with respect to the treatment of women with members of an Islamic society, where women are routinely subjugated to the rule of men, because our respective cultural rules with respect to the treatment of women (whatever *those* are) are not justified by further reasons (Haidt & Bjorklund, 2008b, pp. 215-216). Consequently if someone asks a series of "why" questions with respect to a person's moral judgment, the answers will ultimately end with an accepted cultural rule for which no further reasons can be given. They are psychologically basic. Moreover, because cultural rules are the ultimate psychological cause of one's moral judgments, and cultural rules are not themselves based in reasons or well worked out arguments, then moral judgments are not, and cannot be, justified either.

Schematizing the Social Intuitionists' argument shows that it is a variant of the Regress Argument:

[4] There are five pairs of innate emotional dispositions, according to Social Intuitionists: harm/care, fairness/reciprocity, authority/respect, purity/sanctity, in-group/out-group. Social Intuitionists argue that our dispositions to have emotional responses to this "set of concerns" is innate, and for which, they suggest, a specific evolutionary story could be told (Haidt & Bjorklund, 2008a, p. 203).

1. The reasons for a moral judgment terminate in culturally accepted rules.
2. Culturally accepted rules are not based in reasons.
3. Therefore, moral judgments are not ultimately based in reasons.
4. Therefore, moral judgments ultimately cannot be justified by reasons.

Thus, both Social Intuitionists and Constructive Sentimentalists argue that the reasons a person can offer for their moral judgments terminate in some basic moral claim for which no further reasons can be given. Constructive Sentimentalists identify these basic moral claims as grounding norms (sentiments), whereas Social Intuitionists identify them with accepted cultural rules.[5] Irrespective of this difference their point is the same: moral judgments cannot be justified by reasons because, ultimately, they are not based in reasons.

If the Regress Argument is sound, then it does succeed in drawing an epistemic conclusion from a set of purely psychological claims, and it is possible, then, to derive an "ought" directly from an "is," in the sense that the applicability of some "ought" claims can be ruled out by a set of psychological "is" claims. However, there are two very serious problems with the Regress Argument. The first is that the argument involves an equivocation between two very different senses of a reason, and thus the Regress Argument is invalid. Second, the Regress Argument proves too

[5] This difference in structure also explains why Social Intuitionism implies cultural relativism, while Constructive Sentimentalism implies subjectivism. Social Intuitionists urge that "Moral facts are facts only with respect to a community of human beings that have created them…All ethical statements should be marked with an asterisk, and the asterisk refers down to a statement of the speaker's implicit understanding of human nature as it has developed within his culture" (Haidt & Bjorklund, 2008a, p. 214). Similarly, Prinz holds that Constructive Sentimentalism is straightforwardly subjectivist (Prinz, 2007, pp. 133, 173), because "Each of us is the ultimate arbiter of our own values" (pg. 185). Prinz argues that it is more useful to talk in terms of "communities of moralizers" (pg. 186), presumably because morality can only function as morality if it is somehow shared. This is the conclusion he wants, but not the one he can get given the structure of his theory.

much, in that it would entail that most (if not all) judgments across a variety of cognitive domains cannot be justified. This provides good reasons to reject the Regress Argument.

2. Reasons: Explanation versus Justification

I shall take Prinz's Regress Argument as my main target in this section, though the points I shall make apply equally well to the argument offered by Social Intuitionists. Prinz's Regress Argument is:

1. The reasons for a moral judgment terminate in a grounding norm.
2. Grounding norms are not based in reasons.
3. Therefore, moral judgments are not ultimately based in reasons.
4. Therefore, moral judgments ultimately cannot be justified by reasons.

This problem with this argument is that the move from (3) to (4) involves an equivocation between two very different senses of a reason. In (3), "reasons" refers to explanatory reasons, whereas in (4), "reasons" refers to justificatory reasons. The distinction between explanatory and justificatory reasons is an elementary one in the literature on reasons (Finlay & Schroeder, 2008; Smith, 1994, p. 94). An *explanatory reason* is a reason that explains why a person arrived at a particular judgment, belief, or feeling, or acted a certain way. A *justificatory reason* (or normative reason), on the other hand, is a consideration that provides rational support in favor of or against judging, believing, or feeling a certain way or acting a certain way.[6]

[6] Recall that I am using the term reason in a very thin sense as any consideration that can enter into moral thinking. I do not intend to be making any substantive claims with respect to the structure of reasons or the scope of practical reasoning, e.g., whether reasons are intrinsically motivating, or

This distinction is not one of kind—explanatory and justificatory reasons are not distinct *kinds* of reasons—rather, they are answers to two different kinds of questions that we might ask of a person's attitudes, judgments, and actions (Dancy, 2000, p. 2). The first question is what motivated or caused a person to have an attitude or judgment of a certain kind or act a certain way. This is typically a "why" question. Why did Jones get a glass of water? Why is Smith terrified of snakes? The answer to a "why" question is an explanatory reason—it is what explains why a person acted a certain way or has an attitude of a certain kind. The second question is whether there are any *good* reasons for acting a certain way or holding an attitude of a certain kind. This is a normative question, and it asks whether there are any reasons that recommend or rationally support a person's acting a certain way or holding an attitude of a certain kind. This a typically a "should" question: Should Jones get a glass of water? Should Smith be terrified of snakes? The answer to a "should" question is a justificatory reason—it is a reason that rationally recommends the attitude, action, or belief. Though explanatory and justificatory reasons are not distinct kinds of reasons, the distinction still marks an important difference between two different ways the term is used, and thus between two different senses of the word "reason."

Because explanatory and justificatory reasons answer different kinds of questions, what is an explanatory reason might not be a justificatory reason, and *vice versa*. Indeed, there is no principled limit on what can count as an explanatory reason, just so long as it stands in the right explanatory relationship to the action or attitude in

---

whether practical reasoning is ever only instrumental reasoning (see, for example, Harman, 1975; Korsgaard, 1986; Williams, 1981).

question. If someone wants to know why Smith is terrified of snakes it may be because of some childhood trauma, a movie he watched, a neurochemical imbalance, a targeted magnetic field, an emotional manipulation, hypnosis, an overly-strong genetic predisposition to fear snakes that evolved in hominids, a solar flare, or any one of a thousand other possibilities. Any one or some combination of these reasons could explain Smith's terror of snakes, just so long as they stand in the right explanatory relationship to Smith's terror of snakes; i.e., explains why he is terrified of snakes when he sees one.

But it is possible to step back and ask whether Smith should be terrified of snakes. There is, whether there are any *good* reasons that provide rational support or recommend in any way the attitude of being terrified of snakes. And when asking that question it is clear that the reasons that can be offered are of a different sort. None of the possible explanatory reasons given above are good reasons—they do not *justify* the attitude of being terrified of snakes. For Smith to be justified in being terrified of snakes it must be the case that his attitude is "fitting," that it is appropriate or correct to have that attitude towards snakes (D'Arms & Jacobson, 2000a). For snakes to be a fitting object of terror it must be the case that snakes are the sorts of things it is appropriate to feel terror at, which requires considering what reasons there are that recommend the attitude of terror towards snakes, such their being dangerous or the like. Again, this is just the difference between explanations and recommendations.

However, it is possible for a reason that explains a person's attitude toward some object can also be a reason that justifies that person's attitude towards that

object.[7] Coming to believe the conclusion of a sound deductive argument because it is the conclusion of a sound deductive argument is a paradigmatic instance of such overlap. The reasons that explain the person's belief are the same reasons that justify that person's belief. This is partly why it is better to think of explanatory and justificatory reasons as answers to different kinds of questions rather than as distinct kinds of reasons.

The distinction between explanatory and justificatory reasons is crucial in understanding what goes wrong with the Regress Argument. Both Social Intuitionists and Constructive Sentimentalists argue that the *explanatory reasons* people offer for their moral judgments, the answers to a "why" question, terminate in some basic psychological fact (sentiments or accepted cultural rules), and thus that moral judgments are not derived from reasons. And from this they conclude that that there are no further possible *justificatory reasons* for a moral judgment. But it is now clear that this move involves an equivocation. The fact that a moral judgment is caused in a certain way, and is thereby explained in a certain way, does not straightaway show it cannot be justified by other, justificatory reasons. Of course chains of explanatory reasons must terminate somewhere, but one can still ask whether a moral judgment is appropriate, correct, or the moral judgment one should have—that is, it is possible to ask what justificatory reasons there are that rationally support or otherwise recommend acting a certain way or holding an attitude of a certain kind.

[7] On some accounts of reasons, a reason cannot be a justifying reason for a person unless it also can be (or is) an explanatory (or motivating) reason for that person (Dreier, 1990; Harman, 1975; Williams, 1981). This sort of view follows naturally from reasons internalism, the claim that all genuine reasons necessarily motivate (and thereby explain) a person to act of a certain kind or adopt an attitude of a certain kind. This view of reasons is often attributed to Hume, and is sometimes called the Humean theory of reasons.

An example might make this point clearer. Suppose that without your knowing it, you were slipped a pill in the past that caused you to form the belief that Napoleon lost the Battle of Waterloo.[8] Suppose that this was an experiment being performed to test the efficacy of belief-causing pills, and through some class-action lawsuit you just learned that you were one of the people who had been given this pill many years ago. Given that your belief that Napoleon lost the Battle of Waterloo was caused by a pill, and thus could not be explained by being based in, or derived from, appropriate justificatory reasons, would you now think that your belief that Napoleon lost the Battle of Waterloo is beyond rational criticism? Of course not, because it is still possible to ask whether the belief that Napoleon lost the Battle of Waterloo is appropriate, correct, or the belief you should have; that is, it is still possible to ask what justificatory reasons there are for your belief. Indeed, it is perfectly reasonable to suppose that having learned of the causal etiology of your belief, you would undertake just this sort of task by checking the history books or an encyclopedia and considering the reasons and evidence for the claim that Napoleon lost the Battle of Waterloo, which, in fact, he did. The reasons that explain your belief (the pill) have no direct bearing on whether the belief can be rationally supported by the right kind of justificatory reasons, such as the reasons provided by the historical record, the opinion of experts, or the like. Beliefs, regardless of their causal etiology, are rationally criticizable because they can, or can fail to be, rationally supported by

[8] This example is adapted from Joyce (2006, pp. 179-181). Joyce uses this example to argue that an evolutionary explanation for our capacity for moral judging should undermine our confidence that our moral judgments are ever correct. Joyce's argument is more subtle than either Social Intuitionists' or Constructive Sentimentalists' and so the considerations raised in this chapter do not bear on it directly. However, for a rebuttal of recent evolutionary debunking arguments, including Joyce's, see (Wielenberg, 2010).

justificatory reasons that can show whether the belief is appropriate correct, or the one that we should have.

Similarly, regardless of the causal etiology of a moral judgment, it is possible to ask whether there are potentially justificatory reasons that support it. It is not sufficient to establish the conclusion that there are no such justificatory reasons simply by pointing to the fact that the reasons that explain that moral judgment terminate in something psychologically basic. But perhaps Social Intuitionists or Constructive Sentimentalists will argue in response that sentiments and emotions are somehow different. Perhaps sentiments and emotions, unlike beliefs, are not subject to rational criticism because they cannot be true or false, and the categories of true and false are ineliminable from our concept of justification. But this move will not work because sentiments and emotions are generally among the sorts of things that people typically think can be, or fail to be, supported by justificatory reasons. We regularly criticize our own and other's emotions when they do not fit the circumstances or the object towards which they are taken because we recognize that emotions can, or can fail to be, appropriate, correct, or the emotions we should have (D'Arms & Jacobson, 2000a; Gibbard, 1990; Zagzebski, 2003). Rational criticism of our own and other's emotions does not necessarily change those emotions, of course, because emotions are not always directly responsive to justificatory reasons.[9] Emotions can persist even when people recognize that they are inappropriate. People with phobias, in fact, are often in this sort of situation, where they recognize that their

[9] Neither are beliefs, for that matter. People cannot simply choose to believe something, regardless of how well supported they think it is.

own emotions are inappropriate, or more strongly, *irrational*, but they persist nonetheless.[10]

The first problem with the Regress Argument, then, is that there is no straightforward move from certain psychological claims that moral judgments are caused by sentiments or emotions (grounding norms or accepted cultural rules respectively) to the normative conclusion that moral judgments are not and cannot be justified.[11] Such a move involves an equivocation between two very different senses of the term "reason."

The second problem with the Regress Argument is that it proves too much, for it is certainly the case that many judgments, across a variety of cognitive domains, terminate in something that is psychologically basic and yet this fact does not call into question whether such judgments can be justified. For example, judgments with respect to geography terminate in something psychologically basic, namely, one's beliefs regarding geography. Few, if any, people's beliefs regarding geography terminate in well worked out arguments or are based in further reasons. A person's judgment that Canada is north of Australia is based in that person's beliefs with respect to the relative positions of Canada and Australia, and the general reliability of maps, and these beliefs are psychologically basic in the sense that there are not derived from further reasons or well worked out arguments. But the fact that geography judgments terminate in something psychologically basic does not

[10] A mental state does not have to be reasons-responsive (responsive to justificatory reasons) to be rationally justified, as some claim (Kennett, 2006; Kennett & Fine, 2009). Agents, not mental states, are reasons-responsive, if anything is.

[11] Both Kamm and Berker make similar points with respect to current fMRI research (Berker, 2009; Kamm, 2009). The point is entirely general: no normative conclusions follow *directly* from psychological or neurological findings.

somehow challenge the possibility that some geographical judgments are justified, while others are not. The Regress Argument, then, does not point to a special problem for moral judgments at all; it points to a very generally fact about human judgments, namely, that they terminate in something psychologically basic.

But perhaps Social Intuitionists and Constructive Sentimentalists have a way of arguing that there really is a special problem here for moral judgments. They could argue that beliefs are formed by some set of mechanisms that are intended to track the truth. There is a fact of the matter with respect to the relative positions of Canada and Australia, and belief forming mechanisms operate well when they form true beliefs with respect to the relative positions of Canada and Australia. On the other hand, there is no fact of the matter when it comes to grounding norms, as evidenced by the moral diversity that exists among different cultures.[12] Thus, the mechanisms that form grounding norms or accepted cultural rules are not meant to track the moral truth, but rather, to form those grounding norms or accepted cultural rules that will cause moral judgments that conform to the moral judgments of others in one's culture (Haidt & Bjorklund, 2008a, pp. 206-210; Prinz, 2007, pp. 183-187). Therefore, there is a special problem with respect to the justifiability of grounding norms or accepted cultural rules than there is for beliefs, because they are formed by mechanisms that are not truth tracking.

However, this move will not help to avoid the problem that the Regress Argument proves too much, because even if geography judgments, for example, terminate in beliefs that track the truth, it is still the case that such judgments

[12] Both Social Intuitionists and Constructive Sentimentalists point to the fact of moral diversity between cultures as evidence for metaethical relativism—that the appropriateness of a moral judgment is determined by one's culture (Haidt & Bjorklund, 2008a, pp. 214-215; Prinz, 2007, pp. 254-312)

terminate in beliefs for which a person can give no further consciously accessible reasons. Even if the belief that one's geography beliefs are derived from processes that aim to track the truth is a one of the reasons a person can give for their geography judgments, few people will be able to provide reasons for that belief, or a well-worked out argument on its behalf. Thus, for most people, geography judgments terminate in some belief for which no further reasons can be given, which, according to the Regress Argument, entails that such judgments cannot be justified. So, if the Regress Argument were applied generally to human judgments in a variety of cognitive domains, it would lead to skepticism in most, if not all, of them. This indicates that the real problem here is not with moral judgments, but with the Regress Argument that is somehow meant to undermine them.

3. Psychology and Epistemology

Though the Regress Argument ultimately fails, Social Intuitionists and Constructive Sentimentalists are right in thinking that the underlying psychology of moral judging can illuminate the epistemic project in ethics in some important ways—they are just wrong in arguing that their models of moral judging have direct implications with respect to the justifiability of moral judgments. Regardless, there is an important question here, namely, to what extent do psychological claims illuminate the epistemic project in ethics? That is, in what ways does the psychology of moral judging bear on questions in moral epistemology?

There are deep and interesting issues here (see, for example, Held, 1996), but one important way that the psychology of moral judging can illuminate the epistemological project in ethics is that models of moral judging can show what sorts

of moral decision procedures or processes are not possible for creatures like us given the underlying psychology of our capacities for moral judging and moral reasoning. Such findings can have important implications with respect to epistemic project in ethics. If, for example, moral justification requires that people reason in a certain way, or employ a decision procedure of a certain kind, and the underlying psychology of our capacities for moral judging and moral reasoning reveals that such reasoning is not possible for creatures like us with respect to moral questions, then it follows that moral judgments cannot be justified.

It is this line of reasoning that really seems to be motivating Social Intuitionists and Constructive Sentimentalists skepticism with respect to the justifiability of moral judgments, but here again their target is too narrowly focused on the deductive model of moral judgment. Recall that Social Intuitionists and Constructive Sentimentalist take it that current empirical findings show that the deductive model of moral judgment is wrong. According to the deductive model of moral judgment, moral judging is a process of deduction from first principles of morality. However, the deductive model can be interpreted as both a psychological model of moral judging, and an epistemological model that specifies the reasoning process that is necessary to justify a moral judgment: a moral judgment is justified just in case it is a sound deduction from first principles of morality. If this is the right account of moral justification, and is not possible for creatures like us (or only very rarely possible) to reason deductively to a moral judgment, as Social Intuitionists and Constructive Sentimentalists maintain, then it follows that moral judgments cannot be (or are only very rarely) justified.

This is a much stronger argument than the Regress Argument because it does not attempt to draw an epistemic conclusion directly from some particular psychological claim; rather, it draws an epistemic conclusion from both a psychological and an epistemic premise. And importantly, the epistemic premise in this argument that leads to the skeptical conclusion that Social Intuitionists and Constructive Sentimentalists want, but that the Regress Argument cannot establish, is that for a moral judgment to be justified it must be the conclusion of a sound deduction from first principles of morality. Moreover, it is clear from the structure of the Regress Argument that both Social Intuitionists and Constructive Sentimentalists assume this picture of moral justification, because what the Regress Argument establishes is that moral judgments are not, or are only rarely, conclusions of sound deductions from first principles of morality, and are thus not based in reasons in the way required by the deductive model of moral judgment.

Social Intuitionists and Constructive Sentimentalists are right in arguing that people rarely engage in the sort of deductive reasoning as envisioned by the deductive model of moral judgment, but even so, Social Intuitionists and Constructive Sentimentalists have not established their skeptical conclusion that moral judgments cannot be justified, because neither provide any reason for thinking that the deductive model of moral judgment provides the right account of the requirements of moral justification. This is an important oversight, because their skeptical conclusions can be easily avoided simply by rejecting the deductive model of moral judgment as the correct model of moral epistemology. Indeed, this precisely what we should do.

Rejecting the deductive model of moral judgment as the correct theory of moral epistemology is not an *ad hoc* attempt to vindicate the justifiability of moral judgments; it is entirely consistent with a plausible methodological principle in moral psychology, namely, Flanagan's *Principle of Minimal Psychological Realism* (Flanagan, 1991). It states:

> Principle of Minimal Psychological Realism (PMPR): Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior described are possible, or are perceived to be possible for creatures like us (Flanagan, 1991, p. 32).

According to the Principle of Minimal Psychological Realism, the psychology of moral judging can illuminate the epistemological project in ethics in one further way; namely, if anyone is to recommend a certain moral theory or theory of moral justification, then satisfying the requirements of that theory cannot require a people to do more, from a psychological point of view, then it is actually possible for us to accomplish. This constraint is simply a psychological application of the voluntarist principle that "ought implies can" (Flanagan, 1991, p. 340*n*1). If people ought to judge consistently with the requirements of some moral theory or decision procedure, then it must be the case that people can judge or reason (or at least be thought able to do so) consistently with those requirements.

If it is case that people only rarely, if ever, reason consistently with the requirements of the deductive model of moral judging, then that gives us reason to

reject the deductive model of moral judging as the correct account of moral epistemology, because the justifiability of moral judgments is an ineliminable aspect of ordinary moral experience (Held, 1996; Horgan & Timmons, 2007; Strawson, 1962). As Held writes:

> Moral experience finds us deliberating about which moral recommendations to make to make into our reasons for acting, and reflecting on whether, after acting, we consider what we have done to be justifiable. It finds us weighing the arguments for evaluating the actions of others one way or another, and evaluating the states of affairs we and others are in and can bring about. It finds us approving or disapproving of the traits and practices we and others develop and display (pp. 72-73).

Of course, our introspective experiences of these aspects of moral practice do not provide insight into the underlying processes of moral judging, but an account of moral reasoning and moral justification that is consistent with these aspects of moral experience is preferable to one that eliminates them, as Social Intuitionism and Constructive Sentimentalism do. Indeed, both Social Intuitionism and Constructive Sentimentalism recommend a sort of quietism with respect to some of these aspects of moral experience. For example, if moral judgments are caused by psychologically basic grounding norms or accepted cultural rules, for which no rational dispute is possible, then the proper response to moral disagreements with respect to the traits and practices of others is not the fist, but a shrug of the shoulders. For example, if

people in Islamic societies truly occupy a different moral universe, as Social Intuitionists claim (and Constructive Sentimentalists should agree), with incommensurable moral values, then we in the West literally cannot morally engage with such people or evaluate their traits and practices. The same holds for those in our own society with whom we have incommensurable moral values. If this is truly the case, then quietism with respect to morality is the only appropriate response. And notice, this disconnects morality from the very things we are supposed to care about, namely, how other people think and behave, and thus it fails to capture important aspects of moral experience.

To be clear, it is not the case that a model of moral judging must *vindicate* all aspects of ordinary moral experience. Empirical and philosophical work may sometimes provide good reason for thinking that some aspects of our ordinary first-person moral experience are ungrounded or unwarranted. That much is to be allowed, but the point here is that one criterion of choice between competing models of moral judging is that, *ceteris paribus*, a model of moral judging and its account of moral reasoning is preferable if it is consistent with our ordinary moral experience. Therefore, if there is a highly plausible alternative account of the requirements of moral justification, and an account of moral reasoning that shows how it is possible for creatures like us with minds like ours to satisfy those requirements, then there is reason to prefer both this account of moral justification and the account of naturalized moral reasoning I developed in the previous chapter. As it turns out, there is such a highly plausible alternative account of the requirements of moral justification that it is possible for creatures like us, with minds like ours, to satisfy consistent with the

naturalized account of moral reasoning I developed in the previous chapter, namely, reflective equilibrium.

4. Reflective Equilibrium[13]

Social Intuitionists and Constructive Sentimentalists are right to attempt to situate the epistemological project in ethics within the constraints of the psychological processes of moral judging. This is an important part of naturalizing moral judging and moral reasoning. My aim in this section is to show that there is a better way to do this by showing that there is a fairly influential account of moral justification that is consistent with the picture of naturalized moral reasoning that I developed in the previous chapter. Therefore, there is a fairly straightforward account of moral justification that is possible for creatures like us, and therefore consistent with the requirements of the Principle of Minimal Psychological Realism.

As I argued in the previous chapter, moral reasoning takes place within the particular point of view an agent with particular experiences, intuitions, beliefs, and some (possibly inchoate) moral-theoretic considerations, and consists in subjecting one's moral attitudes, commitments, and judgments to reflective scrutiny to achieve broad coherence among them and with one's experiences, intuitions, and non-moral beliefs; and then modifying, revising, or rejecting some of one's moral attitudes, commitments, and judgments in order to achieve (greater) coherence. In moral reasoning one can consider arguments offered by others, but they will be evaluated against the backdrop of that particular person's experiences, intuitions, and non-moral beliefs.

[13] I am indebted to Ryan Fanselow for helpful discussions of reflective equilibrium.

This view of moral reasoning is entirely consistent with the requirements of moral justification as given by the method of reflective equilibrium. John Rawls introduced the method of reflective equilibrium in his, *A Theory of Justice* (Rawls, 1999).[14] There are two possible interpretations of how the method of reflective equilibrium proceeds: narrow and wide. On the narrow interpretation, the method of reflective equilibrium begins with a person's considered moral judgments. These consist of a person's moral intuitions, excluding those that are formed under circumstances where a person is likely to err.[15] These include those intuitions "made with hesitation, or in which we have little confidence. Similarly, those given when we are upset or frightened or when we stand to gain one way or another can be left aside" (Rawls, 1999, p. 47). The next step in the process is for the person to posit a set of principles that systemize their considered moral judgments, such that the principles posited yield only the considered moral judgments the person already has. If such a set of principles is to be found, the process is complete. However, more likely there will be some inconsistency between the person's considered moral judgments and moral principles, in which case the person has a choice between modifying their considered moral judgments to accommodate their moral principles, or modifying their moral principles to accommodate their considered moral judgments. Once one achieves a consistent set of considered moral judgments and moral principles, one has achieved narrow reflective equilibrium.

[14] I am using the revised edition of *A Theory of Justice*. The first edition was originally published in 1971.

[15] According to Rawls, a moral intuition is a moral judgment that is not " determined by a conscious application of principles so far as this may be evidenced by introspection" (Rawls, 1951, p. 183).

Wide reflective equilibrium is similar in structure to narrow reflective equilibrium, except one adds to the set of considerations taken into the process alternative sets of moral principles, and the relevant arguments for them. Daniels refers to this third set of considerations as background theories (Daniels, 1979). This imposes a further constraint on the process of reflective equilibrium, and is meant to ensure that the moral principles settled on in method of reflective equilibrium are consistent with a broad range of a person's moral and non-moral beliefs. As in narrow reflective equilibrium, the process of wide reflective equilibrium involves a back and forth among one's considered moral judgments, moral principles, and background theories until one arrives at a consistent set. When a person has a consistent set of considered moral judgments, moral principles, and background theories, he or she has achieved wide reflective equilibrium.

Rawls prefers the method of wide reflective equilibrium to that of narrow, because it subjects one's moral intuitions to greater scrutiny. In what follows I shall understand reflective equilibrium to be wide reflective equilibrium. But, one question that still needs to be addressed is why should one think that the process of wide reflective equilibrium is one that justifies a person's moral judgments. To begin, as I argued in the previous chapter, moral reflection can only take place from within the point of view of a particular agent with particular experiences, intuitions, beliefs, and some (possibly inchoate) moral-theoretic considerations. As Tiberius puts it: "The commitments we endorse in reflection are not chosen *ex nihilo*; we must have commitments in the first place, in order to have a reflective point of view on them" (Tiberius, 2008, p. 67, emphasis in original). That is, it is not possible (nor desirable)

to build a set of moral commitments, attitudes, and judgments from the ground up—reflection starts somewhere, and it is within the point of view of a particular person, with a particular history, upbringing, experiences, and the like. Reasoning is not about attempting to take the "view from nowhere," but working out the view of oneself from the inside.

Because moral reasoning begins from within the point of view of a particular person, the right question to ask with respect to moral justification is under what conditions is it appropriate for a person to hold a particular moral attitude, commitment, or judgment. And the answer on this account is that it is appropriate just in case that particular moral attitude, commitment, or judgment hangs together in the right way with the person's other moral attitudes, commitments, judgments, moral-theoretic considerations, and background theories. As Rawls writes: "[Rational] justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view." (Rawls, 1999, p. 502).

Reflective equilibrium as a theory of moral justification is quite possible for creatures like us, with the capacities that we actually possess. Moral reasoning consists in subjecting one's own moral commitments, attitudes, and judgments to reflective scrutiny from within one's own point of view with one's own particular history, experiences, and upbringing. And because this is what moral reasoning consists in, even though many moral commitments, attitudes, and judgments do not terminate in well-thought out arguments, that hardly implies that it is not possible to reflect on them in light of one's other moral commitments, attitudes, and judgments.

Because the process of reflective equilibrium takes as its starting points a person's own extant moral commitments, attitudes, and judgments, it is limited in some important ways. First, some of a person's moral commitments, attitudes, and judgments may be more or less fixed points in their moral thinking. Some of these moral commitments, attitude, and judgments may be the result of cultural upbringing, or they may reflect biological constraints on the human mind/brain (Flanagan, 1991). If the latter is the case, it would help explain the cross-cultural universality of some moral judgments, such as in the Trolley case (O'Neill & Petrinovich, 1998), and it might imply that there are a fixed number of possible human moralities as some moral psychologists claim (Dwyer, 2006, 2009; Hauser, 2006). Regardless, because some of a person's moral commitments, attitudes, and judgments are more or less fixed points in their moral thinking, reflective equilibrium is conservative, in the sense that it does not generally lead to a radical revisioning of one's moral commitments, attitudes, and judgments.

Second, even when a person decides that some moral commitment, attitude, or judgment should be accepted, rejected, or revised, it is not always the case that their moral commitments, attitudes, and judgments are directly corrigible by their conscious reasoning and decision. A racist who comes to the view that his own racist attitudes are wrong, and desires to change them, cannot simply, by so deciding come to change his underlying attitudes. Instead, he may find that he needs to override his initial racist judgments in light of his explicit moral commitments, attitudes, and theories. With practice, such overriding could become reflexive. Or, such a person could adopt some strategies for indirectly changing his underlying racist attitudes,

such as reading stories where minorities are figured prominently as protagonists. What this reveals is that when a person cannot directly change his or her underlying moral commitments, attitudes, or judgments, he or she can adopt a meta-attitude with respect to how to treat that moral commitment, attitude, or judgment in thinking or acting. As I said before, it is not, in the first instance, mental states that are reasons responsive, but agents. An agent may be reasons responsive without some of her mental states being directly reasons responsive. However, such reflexive and habitual overriding can still be connected to reasons in the right way, because it is a rational strategy adopted by an agent in response to a recalcitrant mental state.

Third, moral reasoning may not be able to settle all moral disagreements decisively. Because cognitively limited creatures such as ourselves must take their own moral commitments, attitudes, and judgments as the starting points for moral reasoning, disagreements can easily arise between people who have different starting points and no amount of shared moral reasoning can decisively settle the issue between them. Even when there is wide agreement between people with respect to their deliberative starting points, it is not the case the moral reasoning, when done well, recommends or requires a single right answer to some difficult moral questions. This is an unavoidable conclusion if one takes seriously our cognitive limitations, but it is not necessarily a skeptical one. Scanlon (2000) draws a useful analogy with scientific judgments and moral judgments on this point. He writes:

> Disagreement about which of several competing scientific hypotheses is best supported by the available evidence, for example, often persist even among inquirers who are experts in the field. Further evidence

> may determine which of these hypotheses was correct, but disagreement about reasons—about which hypothesis the more limited body of evidence in fact supported—may continue, especially when the inquirers are committed to different scientific of methodological programs. Persistent disagreements about right and wrong have a similar character: they are disagreements about how complex sets of conflicting reasons should be understood and reconciled, and they are most likely to persist when people's differing interests and commitments lead them, in different ways, to concentrate on certain of these reasons (and on certain ways of understanding them) and to neglect others (pg. 358).

Regardless of these three limitations, the process of reflective equilibrium can justify a person's moral commitments, attitudes, and judgments if they cohere in the right way, providing justificatory support, that is, justificatory reasons, for each other. Through such a process of reasoning a person can discover that some of his or her moral commitments, attitudes, and judgments are not justified, and take steps to change them. Moreover, this account of moral justification and moral reasoning is consistent with our ordinary moral experiences, and is therefore preferable to the accounts of moral reasoning and moral justification given by Social Intuitionists and Constructive Sentimentalists.

5. Conclusion

Over the past two chapters, I have defended a particular view of what moral reasoning consists in, and I have argued that it provides a much better account of the facts of moral change, that it can accommodate central features of our moral experience, and that it is psychologically possible for creatures like us. But Social Intuitionists and Constructive Sentimentalists will take issue with this last point, because they argue that current empirical findings with respect to the psychology of moral judging show that emotions are the real causes of moral judgments, and thus that moral reasoning has little or no role in moral judging or in our broader moral psychology. That is, they argue that the fact that emotions have an important causal role in moral judging rules out the possibility that moral reasoning can also have an important causal role in moral judging. In the next chapter I shall identify an underlying assumption in this line of reasoning, and then argue that the assumption ought to be rejected. By rejecting this assumption, it will be possible to develop a framework for the psychology of moral judging that avoids the theoretical and explanatory problems of Social Intuitionism and Constructive Sentimentalism, and show how both the emotions and reasoning can have important causal roles in moral judging, and in our broader moral psychology.

Chapter 5: Unity and Disunity

It is now possible to return to the first question at the end of the second chapter: does recognizing that emotions play a central causal role in moral judging require giving up a central causal role for distinctively moral reasoning? Again, the answer is "no," but why do Social Intuitionists and Constructive Sentimentalists think that it does? My diagnosis is that they are both implicitly committed to the Unity of Process Thesis, which is the claim that all *genuine* moral judgments are the products of a single "core" psychological process. However, I shall argue that under the constraints of the Unity of Process Thesis it is not possible to provide an adequate model of moral judging. Drawing on dual-process accounts of reasoning and judgment, it is possible to move beyond the Unity of Process Thesis, and develop a dual-process account of moral judging that distinguishes between two kinds of moral judgment, intuitive and deliberative, that are subserved by different kinds of psychological processes. Developing a framework for moral judging and judgment that distinguishes between intuitive and deliberative moral judging and judgments makes it possible to capture important features of our moral psychology that neither Social Intuitionism nor Constructive Sentimentalism can, and is, moreover, consistent with current accounts of the general architecture of human judgment.

1. The Unity of Process Thesis

The Unity of Process Thesis is the claim that all *genuine* moral judgments are the products of a single "core" psychological process, and it is the implicit acceptance of this thesis that makes it seem as though recognizing that emotions have an

important causal role in moral judging requires rejecting an important causal role for reasoning.[1] Under the constraints of the Unity of Process Thesis theorizing with respect to the underlying processes of moral judging is cast in terms of an opposition between reasoning and emotions: Moral judging is explained by one, or the other; but not both. Thus, if emotions are the "core" psychological process of moral judging, then under the Unity of Process Thesis, reasoning can have only some peripheral causal role in moral judging, such as determining whether some object of appraisal falls under a particular concept.

Moreover, it is not simply new wave sentimentalists who implicitly accept the Unity of Process Thesis; their critics do as well (Fine, 2006; Horgan & Timmons, 2007; Jones, 2006; Kennett, 2006; Kennett & Fine, 2009). Indeed, the Unity of Process Thesis provides the best way of understanding the ongoing debate between new wave sentimentalist and their critics, because, at its heart, the debate between new wave sentimentalists and their critics revolves around determining which process is the "core" psychological process of moral judging, reasoning or emotions.

Once one recognizes that the Unity of Process Thesis is operating in the background of this debate, it helps makes sense of the two general strategies employed by new wave sentimentalists and their critics that follow directly from it. The first strategy involves starting from a particular metaethical claim with respect to what constitutes the essential feature or features of all *genuine* moral judgments, and then moving to a psychological conclusion with respect to what "core" psychological

[1] The term "core" psychological process in this context means only a discrete psychological process that is necessary to moral judging. Some psychologists, notably Spelke, use the term "core" psychological process to refer specifically to those processes that are encapsulated (informationally independent from other cognitive systems), domain-specific, and task-specific (Spelke, 2000). I do not intend this more substantive sense of the term "core" psychological process.

process could give rise to that essential feature or features. For example, it may be that genuine moral judgments involve the emotions, or are based in reasons in the right kind of way, and any putative moral judgment that fails to have this feature is not really a genuine moral judgment after all. Using this strategy, it is then an easy to specify the "core" psychological process if one accepts the Unity of Process Thesis—it is the one the produces whatever the genuine moral judgments are: either reasoning or the emotions.

The second strategy involves first specifying what the "core" psychological process (either reasoning or the emotions) is, and then showing how that "core" process it is ultimately causally implicated in all moral judgments, even those that are (or seem to be) proximately caused in some other way. This requires showing, for example, that reasoning is ultimately involved in all episodes of moral judging, even those that are (or seem to be) proximately caused by the emotions; or that emotions are involved in all episodes of moral judging, even those that are (or seem to be) proximately caused by reasoning. Thus, even though there are some moral judgments that seem to be caused in other ways, there is really just one "core" psychological process after all. These two general strategies lead to four specific positions in attempting to explain moral judging under the constraints of the Unity of Process Thesis, summarized in Table 5.1.

	"Core" psychological process is reasoning	**"Core" psychological process is emotions**
Metaethical Strategy	(1) "Moral" judgments caused by the emotions are not genuine moral judgments	(2) "Moral" judgments arrived at solely through reasoning are not genuine moral judgments
Psychological Strategy	(3) All moral judgments are ultimately products of moral reasoning, even those that are proximately caused by the emotions	(4) All moral judgments are ultimately products of the emotions, even those that are proximately caused by reasoning

Table 5.1 summarizes the dialectic with respect to judgments.

The four options line up with the current positions taken in the ongoing debate between new wave sentimentalists and their critics:

(1) Only moral judgments that are the products of reasoning are genuine moral judgments; "moral" judgments caused by the emotions are not genuine moral judgments. Some prominent critics of new wave sentimentalist models of moral judging defend this option, including Jones (Jones, 2006) and Kennett and Fine (Kennett & Fine, 2009). According to these philosophers, it is a conceptual truth that genuine moral judgments are derived from moral principles in the right sort of way, and based on reasons that are accessible to all.

(2) Only moral judgments that are the products of the emotions are genuine moral judgments; "moral" judgments caused by reasoning are not genuine moral judgments. This is the option defended by Constructive Sentimentalists (Prinz, 2007), who hold that it is a conceptual truth that emotions are necessary and sufficient for moral judgment.

(3) All moral judgments are ultimately products of moral reasoning, even those that are proximately caused by the emotions. Jones (Jones, 2006) also defends this option, as do Horgan and Timmons (Horgan & Timmons, 2007), who hold that appropriate emotional responses are properly guided by prior episodes moral reasoning.

(4) All moral judgments are ultimately products of the emotions, even those that are proximately caused by reasoning. This is the option is defended by Social Intuitionists (Haidt, 2001; Haidt & Bjorklund, 2008a), who hold that, even when the moral judgment is proximately produced by reasoning, the emotions are the real "driving force" behind moral reasoning, (Greene & Haidt, 2002, p. 522).

It is important to note that the motivation for these various options is not the Unity of Process Thesis itself. Moral psychologists are not trying to make their models of moral judging consistent with the Unity of Process Thesis—it as an implicit assumption behind them, which helps explain the way the contemporary debate with respect to moral judging has been set up. The problem, however, is that each of the four options listed above raises its own set of theoretical and explanatory

worries. It is simply not possible to fit all features of moral judging and judgment into the same explanatory box, as I shall argue. In short, the problem is with the Unity of Process Thesis, and the way the debate has been set up. Making progress requires abandoning the Unity of Process Thesis, but seeing that requires first seeing how the Unity of Process Thesis informs and constrains the current debate.

2. New Wave Sentimentalism

New wave sentimentalists claim that the emotions are the "core" psychological process of moral judging, and under the constraints of the Unity of Process Thesis they argue either that: (1) reasoning does not produce *genuine* moral judgments; or (2) all moral judgments are ultimately products of the emotions, even those that are proximately caused by reasoning. Constructive Sentimentalists defend the first option; Social Intuitionists defend the latter. According to Constructive Sentimentalists, emotions are necessary and sufficient for tokening a moral concept, and so any process that does not produce or token emotions does not and cannot produce genuine moral judgments. Reasoning, according to Constructive Sentimentalists, does not produce or token emotions because it involves the manipulation of affect-free propositional attitudes, such as beliefs, and therefore it cannot produce genuine moral judgments because it can never produce token-instances of moral concepts. Reasoning can produce a string of words that could be used to express a moral judgment, but that string of words itself is not a genuine moral judgment. It is what Prinz calls a "verbalized moral judgment" (Prinz, 2007, p. 100), and the function of these verbalized moral judgments is not to express a genuine moral judgment, but to reason about moral values a person do not actually hold. Prinz

writes:

> We often talk as if verbalizations of moral judgments were moral judgments in their own right [i.e., genuine moral judgments]. I will refer to sentences such as "Pickpocketing is wrong" as a verbalized moral judgment. But this label is really shorthand. "Pickpocketing is wrong" is not a judgment; it's a string of words. In calling it a verbalized moral judgment, I mean it is a verbal form that might be used to express a moral judgment. Verbalized moral judgments are very useful because they allow us to reason about moral values that we don't actually hold (Prinz, 2007, p. 100).

According to Constructive Sentimentalists, it is a conceptual truth that certain emotions are necessary and sufficient for a moral judgment,[2] and thus that reasoning cannot produce genuine moral judgments. This is one way to develop a sentimentalist model of moral judging consistent with the Unity of Process Thesis. Social Intuitionists, on the other hand, defend the second option and argue that all moral judgments are caused by the emotions, even when they seem to be caused by reasoning, or are proximately caused by reasoning. According to Social Intuitionists, moral reasoning generally consists in providing post hoc justifications for one's already arrived at moral judgments. Providing these post hoc justification is an important social skill, because in social contexts we demand of each other reasons for

[2] Prinz argues, borrowing from Rozin et al. (1999), that only those emotions that are part of the CAD triad (contempt, anger, and disgust) are necessary and sufficient for a genuine moral judgment.

our moral judgments (particularly when disagreement is involved), and providing plausible-sounding reasons for one's moral judgments is necessary to deflect criticism, gain social allies, and signal that one is a trustworthy social partner. So even though it appears as though a person's moral judgments are based in reasoning, for the most part the reasons a person offers are simply post hoc justifications that are not the actual basis of the person's moral judgment.

These post hoc reasons, however, can influence another person's emotions in such a way as to cause a new and different moral intuition leading to a new and different moral judgment. The post hoc reasons a person comes up with may cause someone else to see an action or person in a different way, which can trigger a new and different moral intuition, leading to a new and different moral judgment. This can also happen in private reflection, but more often it occurs in social contexts when two or more people are exchanging post hoc reasons for their (possibly divergent) moral judgments. According to Social Intuitionists, the back-and-forth exchange of post hoc reasons constitutes moral reasoning, and such reasoning can play an important causal role in moral judging, but only when it triggers a new emotional response. Thus, all moral judgments, even those proximately or apparently caused by reasoning are still directly caused by the emotions.

Thus, the two most well-developed and influential new wave sentimentalist models of moral judgment align perfectly with the two options available under the constraints of the Unity of Process Thesis, which strongly suggests that it is an implicit assumption motivating these models of moral judging. Under this constraint there can only be one "core" psychological process of moral judging, and the

theoretical space is constrained by a supposed sharp dichotomous choice between reasoning and emotions. However, as I have been arguing throughout this dissertation, both Social Intuitionism and Constructive Sentimentalism raise serious theoretical and explanatory worries with respect to broader aspects of our moral psychology, including moral change and moral justification. These worries are not decisive criticisms of either the Social Intuitionist or Constructive Sentimentalist models of moral judging, but they do show that there are serious problems in trying to account for all of the features of moral judging and judgment and its role in our broader moral psychology with the emotions as the single "core" psychological process. If emotions cannot account for all the features of moral judging and judgment and its role in our broader moral psychology, then one natural approach under the constraints of the Unity of Process Thesis is to argue that reason is the single "core" process of moral judging. This is the approach that critics of new wave sentimentalists take, but as I shall argue, this approach is not satisfactory either.

3. The Critics of New Wave Sentimentalism

Critics of new wave sentimentalism stress the role of reasoning in moral judging, and under the constraints of the Unity of Process Thesis argue that either: (1) new wave sentimentalist accounts ignore the conceptual link between moral judgments and reasons; or (2) that the empirical literature is actually consistent with the claim that all moral judgments are ultimately explained by reasoning. Some prominent critics of new wave sentimentalist models of moral judging take the first approach, and argue that "real" moral judgments must be derived from reasons in the right sort of way because only such judgments have the normative authority to be

properly action-guiding judgments for the individual who makes them (Jones, 2006; Kennett & Fine, 2009). They take this to be a conceptual truth about the nature of genuine moral judgments; a view, they argue, is shared by many moral philosophers. For example, Smith writes: "It is a conceptual truth that claims about what we are morally required to do are claims about our reasons" (Smith, 1994, p. 84). [3] Many other theorists defend similar claims (Deigh, 1995; Svavarsdottir, 1999). Rachels puts the point this way:

> If someone tells you that a certain action would be wrong…you might ask why it would be wrong and if there is no satisfactory answer you may reject that advice as unfounded. In this way, moral judgments are different from mere expressions of personal preferences…[because] moral judgments require backing by reasons, and in the absence of such reasons, they are merely arbitrary. This is a point about the logic of moral judgment…One *must* have reasons or else one is not making a moral judgment at all (Rachels, 1993, p. 483, emphasis in original).

If it is a conceptual or logical truth about moral judgments that they can be backed by reasons in the right kind of way, and current empirical research on so-called moral judgments shows that they are subject to framing effects (Petrinovich & O'Neill, 1996), are influenced by feelings of disgust (Schnall, et al., 2008; Wheatley

[3] Smith here is talking about justificatory reasons, not explanatory reasons. Justificatory reasons are those reasons that can provide rational support for a judgment, while explanatory reasons are simply those that explain why we came to the judgment we did. Some moral rationalists, such as Smith, hold that justificatory reasons must also be explanatory reasons. That is, it must be possible for the reasons that could justify our judgments must also be able to motivate us to act consistently with them (see Williams, 1981 for the locus classicus on this topic). It does not matter for the argument that follows how we understand what constitutes a reason at this point.

& Haidt, 2005) and happiness (Valdesolo & DeSteno, 2006), are affected by strong electromagnetic fields (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010 107, 6753-6758), and that people often cannot give any further plausible reasons for them (Cushman, et al., 2006; Haidt, 2001; Hauser, 2006; Hauser, et al., 2007), then whatever these judgments are, they are not *genuine* moral judgments unless they can be backed by reasons in the right way. They may be moral intuitions, but genuine moral judgments must be backed by reasons in the right way, preferably reasons deriving from a general moral theory (Singer, 2005).[4] Thus, the so-called moral judgments being studied in the laboratory are not clearly *genuine* moral judgments. They are ersatz moral judgments, because they lack the normative authority to be properly action-guiding for the individuals who make them if they cannot be backed by reasons.

This is one possible way to claim that reasoning alone is sufficient to produce a genuine moral judgment, but this sort of argument raises two worries. First, the conceptual claim that genuine moral judgments must be based in reasons in the right way implies that when most people (perhaps including ourselves) judge the permissibility of pulling a lever or pushing a fat man in the trolley and footbridge cases they are not making genuine moral judgments because most people cannot provide any good reasons for their pattern of judgments. Thus, the vast majority of people (up to 90% according to Hauser, 2006) are making ersatz moral judgments when responding these cases. But it is not just these particular cases that should be of concern, because the implication of much of moral dumbfounding research is that

[4] For Singer, a moral intuition is what first comes to mind when presented a case such as the Trolley case. A moral judgment is one's considered judgment with respect to a case, preferably made in accordance with some general moral theory, such as utilitarianism.

most people, most of the time, rely *solely* on quick, initial moral judgments when deciding what should be done, and that none of these judgments are connected to reasons in the right sort of way. If this is right, then this conceptual claim is committed to the highly implausible position that most people, most of the time, are not engaged in a genuine form of morality at all—most people, most of the time, are simply engaged in ersatz morality and making ersatz moral judgments.

This is certainly a legitimate worry, though perhaps those who defend the conceptual link between moral judgments and reasons will reply that most people most of the time can and do produce reasons for their quick, initial moral responses, and moreover, take these reasons to be the rational bases for them, and thus they are genuine moral judgments (Jones, 2006). But this leads to the second worry with this supposed conceptual connection, which is that it does not take the research on moral dumbfounding seriously enough. The phenomenon of moral dumbfounding strongly suggests that the reasons people provide for some of their moral judgments are not the actual basis of those judgments, but are instead post hoc confabulations. These confabulated reasons certainly appear to the person to be the rational basis for their moral judgments, even when they are not. Thus, the reasons people provide for their moral judgments do not provide any reliable guide to the underlying processes of moral judging, and thus provide no support for the claim that most people's moral judgments really are connected to reasons in the right sort of way that is necessary for them to be properly action-guiding.

Second, this supposed conceptual connection between a moral judgment and reasons implies that scientific study has no actual bearing on uncovering the

underlying processes of moral judging. By claiming that the term "moral judgment," as a conceptual matter, applies only to those judgments that are backed in the right way by reasons, these critics of new wave sentimentalism are simply asserting a favored philosophical position that is not empirically falsifiable (or verifiable). The worry here is that this option rules out the possibility that it is at least a partly empirical question what the underlying processes of moral judging are. At most, under this view, it is possible to discover that genuine moral judgment is rare or impossible, but it is not possible to learn what the actual processes of moral judging are through empirical investigation.

These worries are not decisive objections, but they are serious enough to warrant investigating alternative options for showing that reasoning is the "core" psychological process. The second option for critics of new wave sentimentalism is to argue that all moral judgments are ultimately products of reasoning, even when they are proximately the products of emotional responses or some other nonconscious reasoning process (Fine, 2006; Horgan & Timmons, 2007).[5] This would preserve the Unity of Process Thesis by showing that all moral judgments are ultimately the products of the single "core" psychological process of reasoning, even if reasoning is not the proximate cause of all moral judgments. The way to make this argument is to argue that those moral judgments that are not directly deductions from moral principles are nonetheless related to the reasoning in the right way because they are the products of a process of moral experience and education that has its starting point

[5] Fine (2006) defends both options available to critics of new wave sentimentalism: she defends the first option by arguing that there is a conceptual connection between moral judgments and reasons, and defends the second as a means of allaying the worry that the empirical research shows that people rarely, if ever, make genuine moral judgments.

in such reasoning. On this view, some moral judgments are the result of a process that takes moral knowledge-that arrived at through reasoning, and produces moral know-how that can produce moral judgments without having to go through any steps of conscious reasoning (Ryle, 1949). For example, in reasoning a person can come to the conclusion that killing someone for no reason is wrong, and through training and experience, that person can learn to put that deliberative knowledge into practice quickly and intuitively to judge, for example, that this instance of killing is wrong (Herman, 1993). Ideally, the moral judgments of mature moral agents are hard won know-how and expertise in the moral domain—they are evidence of *practical wisdom* that derives from practical reason and experience. Such know-how judgments are still based in reasons in the right kind of way, because they ultimately derive from a process of reasoning.

The picture here is analogous to the way people learn to drive a car. People start by consciously learning the rules of the road and how to manipulate the controls of the car, but through training and experience people simply come to know how to apply those rules and manipulations in all sorts of situations, conditions and permutations without consciously considering what they are doing. And notice too, asking experienced drivers to give their reasons for driving a certain way in a certain situation can lead to a form of driving dumbfounding: they may not be able to give a reason, except perhaps some general rule or rules that do not fully explain their behavior or to cite some intuitive sense of what was going to happen (e.g., "I had a sense that that person was going to swerve into this lane, so I slowed down"). But that

is to be expected: know-how is not consciously accessible propositional knowledge; it is hard-won practical wisdom that manifests itself intuitively.

Similarly, the argument goes, a good deal of the moral judgments of mature moral agents reflects hard-won practical wisdom in the moral domain. They may not know what reasons they have for their moral judgments consciously, but those moral judgments are not therefore suspect, irrational, or anything of the sort, because they are ultimately backed by reasoning, and cultivated through moral training and experience. They reflect the hard-won practical wisdom of mature moral agents—moral know-how that is not consciously accessible and manifests itself intuitively, perhaps even emotionally. This view of moral judging is implicit in the work of virtue ethicists, who stress the need for practical wisdom, usually through exposure to stories or situations that allows people to fine-tune their moral judgments and allows them to gain moral know-how, and for their moral judgments to become "second-nature" (see, for example Lawrence, 1995; McDowell, 1988b). This view is also explicitly developed by a number of moral philosophers, including Herman (Herman, 1993), Gibbard (Gibbard, 1990), and Nichols (Nichols, 2004).

This view raises one serious worry, which is that this view is inconsistent with the developmental trajectory of moral judging in children. Children make quick, intuitive judgments that they recognize as distinctly moral between the ages of 3- and 5-years-old (Turiel, 1983).[6] Moreover, children tend to make some fairly sophisticated moral distinctions starting fairly early as well. For example, 3-4 year-

[6] There is some skepticism with respect to the moral-conventional distinction in the psychological literature (Kelly et al., 2007), but the most natural interpretation of the data cited in this paper doesn't support the authors' conclusion that there is no distinction between moral and conventional transgressions. It is more naturally explained by contextual features.

olds use intent to distinguish morally between two actions with the same outcome; 4-5 year-olds recommend proportional punishments for individuals based on how wrong the action is; and 5-6 year-olds allow the false factual beliefs can be an excusing condition, but not false moral beliefs (Mikhail, 2007). If children are making quick and sophisticated moral judgments at this young age, then, on this view, they must be engaging in some sort of complicated and sophisticated moral reasoning even earlier. But it is not at all plausible to think that children are performing the sophisticated kind of reasoning required to come up with and settle on principles such as the Doctrine of Double Effect, or other sophisticated distinctions. The reasoning envisioned to explain people's quick moral judgments is difficult, even for adults, and so it is quite implausible to think that very young children are engaging in this sort of reasoning to arrive at some sort of know-how. Perhaps *some* moral principles underlying intuitive moral judging are the products of such reasoning, but it is not very likely that it can be true of *all* them.

This worry is not a decisive objection, and it is not intended as such. However, it is quite serious and it reveals that there is a serious problem in attempting to shoehorn all moral judgments into the reasoning box, just as there are serious problems attempting to shoehorn all moral judgments into the emotions box. Ultimately, the problem here is that that whole debate has been set up on the wrong terms by the Unity of Process of Thesis. Under the constraints of this thesis, the debate is stuck as a choice between reason and emotions, but each choice brings with it some serious theoretical or explanatory worries: either by over-intellectualizing the process of moral judging; or by denying that reason has any significant role in moral

judging or in our moral psychology. This is sufficient reason to seriously reconsider the foundations of the debate between new wave sentimentalists and their critics, and the supposed sharp dichotomous choice between reason and emotion at its heart. In short, it is sufficient reason to investigate whether progress can be made by rejecting the Unity of Process Thesis.

4. Dual-Process Cognitive Architecture

Initially rejecting the Unity of Process Thesis may seem like a fairly radical move, but it is actually consistent with a host of empirical research over the past several decades, which indicates that the human mind has a dual-process architecture, and that judgments in many domains can be arrived at by two distinct kinds of mental processes: one that is fast, automatic, and nonconscious; and another that is slow, effortful, and conscious. Processes of the first kind are often called System 1 processes, and processes of the second kind are often called System 2 processes (Evans & Over, 1999; Kahneman, 2003; Kahneman & Frederick, 2002; Sloman, 1996; Stanovich, 1999; Stanovich & West, 2000).[7] In psychology these theories of cognitive architecture are generally referred to as Two Systems or Dual-Process views.

The initial motivation for these views of the architecture of human reasoning and judgment is a number of studies which show that people often make quick and intuitive judgments that conflict with their consciously endorsed norms. For example, Tversky and Kahneman presented subjects with the following vignette, and then

[7] Though they are often simply referred to as System 1 and System 2, but this gives the impression that there are only two systems.

asked them to order eight items with respect to what was most likely to be the case based on the information provided:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student she was deeply concerned with issues of discrimination and social justice and also participated in antinuclear demonstrations (Tversky & Kahneman, 1983).

Among the listed choices were the items, "Linda is a bank teller," and "Linda is a bank teller and active in the feminist movement." Surprisingly, 80% of subjects indicated that Linda was more likely to be a bank teller and a feminist than a bank teller, even though this ordering implies that the conjunction (Linda is a bank teller and a feminist) is more likely than one of its conjuncts (Linda is a bank teller). Tversky and Kahneman labeled this pattern of judgment the *conjunction fallacy*. Even more surprising, however, is that the subjects in these experiments were medical students with statistical training and graduate students in the decision science program at the Stanford Business School, all of whom consciously knew and endorsed the rules governing conjunctions and probabilistic reasoning. It seems, then, that these subjects were not relying on their consciously endorsed rules when making their orderings, but on quick intuitive judgments of similarity. Based on the information given, Linda seems more similar to a bank teller and a feminist than to stereotypical bank teller. These quick similarity judgments lead people to judge that she is more likely to be a bank teller and a feminist than just a bank teller, even

though these probability judgments conflict with their consciously endorsed rules governing conjunctions and probabilistic reasoning.

Moreover, when Tversky and Kahneman pointed out to the subjects in their study that their probability judgments conflicted with the rules of conjunction, most agreed that they had gotten the rank orderings wrong, but others were dumbfounded with respect to their probability judgments in this case—that is, they could not give any reason to support their probability judgments, but they were sure they were right that Linda was more likely to be a bank teller and a feminist than just a bank teller. For example, one subject in Tversky and Kahneman's study attempted to defend his incorrect ranking by saying, "I thought you only asked for my opinion" (Tversky & Kahneman, 1983, p. 300). Stephen Jay Gould, the famous naturalist is more direct in reporting his dumbfounding in this case, "I know that the [conjunction] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me—'but she can't just be a bank teller; read the description'" (quoted in Sloman, 2002, p. 386).

The Linda case is just one example of a situation where people are inclined to make quick intuitive judgments that conflict with their consciously endorsed rules. It is now a robust finding that humans regularly and persistently make intuitive judgments that violate their consciously endorsed rules, even when they have all the resources to reason through a problem carefully and correctly (for a review see Brenner, Koehler, & Rottenstreich, 2002). Many psychologists argue that the potential disparity between intuitive judgments and more carefully reasoned and consciously arrived at judgments cannot be explained by a unified theory of judgment

or a single process theory of cognitive architecture, and that intuitive judgments cannot be explained simply as expertise judgments (Gilbert, 1999). At the very least, two separate and distinct kinds of processes are required to explain human judgment, and dual-process models of the cognitive architecture of human judgment are meant as a way to explain that.[8] System 1-type processes are quick, automatic, and nonconscious (though some of their outputs are conscious); and System 2-type processes that are slow, effortful, and conscious. I use the locutions "System 1-type process" and "System 2-type process" to highlight that dual-process theory is not the claim that there are only two processes in the mind, System 1 and System 2, but that there are two distinct *types* of cognitive processes that operate in importantly different ways (Evans, 2008b).

Beyond this general description, there is little consensus on how best to distinguish between the two types of processes. For example, some theorists have suggested that the two systems differ with respect to evolutionary age, arguing that System 1-type processes are evolutionarily old and shared with other animals, whereas System 2-type processes are evolutionarily recent and distinctively human (Evans & Over, 1996; Greene & Haidt, 2002; Stanovich, 1999). There are good reasons to be skeptical of a cut-and-dried division of the two types of processes in terms of evolutionary age (Evans, 2008a), and it is quite possible that if dual-process architecture is the right way to understand moral judging, that some, if not all, of the processes underlying moral intuitions are evolutionarily recent and distinctively

[8] Moreover, there is evidence that humans have at least two different kinds of memory systems—one kind that is consciously accessible and another kind that is not—and that these different kinds of memory systems can have different contents that can produce conflicting judgments (Bargh, Chen, & Burrows, 1996; Evans & Over, 1996; E. R. Smith & DeCoster, 2000).

human (Hauser, 2006). However, nothing of what follows depends on adopting any particular view about the evolutionary age of the two types of processes.

Another common claim is that System 1-type processes are associative, and System 2-type processes are rule-based (Kahneman & Frederick, 2002; Sloman, 1996, 2002; Strack & Deutsch, 2004). But here too there is good reason for thinking that the characterization of the two types of processes on this basis is too simplistic, and that at least some System 1-type processes are indeed rule-governed (Carruthers, 2009; Evans, 2008a). For example, many researchers claim that System 1-type processes rely on heuristics, which are not merely associative operations, but rather 'rules of thumb' (Evans, 2008a).

Even though there is some disagreement about how best to draw the line between System 1-type processes and System 2-type processes, most psychologists agree that there are at least two different and causally distinct types of systems underlying human thought and judgment; one that is quick, automatic, and intuitive, and the other that is slow, effortful, and conscious. To say that these two types of systems are different and causally distinct does not imply that they do not interact in important ways. For example, System 1-type intuitions can bias System 2-type reasoning in a number of ways. Some intuitive judgments can bias System 2-type reasoning, such that individuals find it more difficult to deviate too much from their initial System 1-type intuition, which can lead to confirmation bias in System 2-type reasoning, where individuals are much more likely to seek out information that

confirms their System 1-type intuition and ignore or downplay evidence that conflicts with it (see Nickerson, 1998 for a review).[9]

System 1-type intuitions can influence System 2-type reasoning, but System 2-type processes cannot influence System 1-type processes in the same way. However, in some cases it is possible for a System 1-type intuition to be overridden by System 2-type reasoning.[10] Indeed, many dual process theorists characterize System 2-type processes in terms of its role in scrutinizing System 1-type intuitions, and either approving or overriding them (Gilbert, 2002; Kahneman & Frederick, 2002; Stanovich & West, 2000).

However, as many studies make clear, people do not frequently engage in System 2-type reasoning, especially if a highly plausible intuition is readily available and relevant to the task at hand. In one such study, Frederick presented students at Princeton and the University of Michigan the following: "A bat and a ball costs $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" (reported in Kahneman, 2003, p. 699). Most subjects reported an immediate tendency to say 10 cents, because the figure $1.10 seems to split naturally in this way. The correct answer, however, is 5 cents, but surprisingly 50% of Princeton students and 56% of University of Michigan students responded that the ball costs 10 cents. These students responded with their initial intuitive judgment, even though they all had the mathematical knowledge to answer the question correctly if they had taken the time

[9] Baron has collected some evidence of what he calls "myside bias" in people's assessments of arguments for and against abortion (Baron, 2003)

[10] Part of the reason System 2-type processes are thought to act as a check on System 1-type processes and not *vice versa* is that System 1-type processes are much quicker than System -type processes. But more importantly, some System 1-type processes are cognitively impenetrable, meaning that they cannot be influenced by the conceptual or intellectual resources of the sort that characterize the operations of System 2-type processes. (The term "cognitively impenetrable comes from (Pylyshyn, 1981), though he does not specifically mention System 1-type or System 2-type processes.)

to reason through it. Those who answered the question correctly, on the other hand, did engage in reasoning, and overturned their initial intuitive judgment of 10 cents.

There are likely many reasons why people often fail to overturn an initial System 1-type intuitive judgment even though they have the time and resources to do so. Some people are simply less likely than others to question their intuitive judgments or to consider evidence that conflicts with them (Stanovich, 2009). Moreover, System 2-type reasoning is easily disrupted by time pressures, multi-tasking, mood, and time of day (Bodenhausen, 1990; Finucane, et al., 2000; Kahneman & Frederick, 2002), and even when people engage in System 2-type reasoning it cannot always override a strong System 1-type intuition (Bargh, et al., 1996; Wilson, Lindsey, & Schooler, 2000). A System 1-type intuition can persist, even when people recognize good reasons to drop or modify it, which is another way of saying that people can be dumbfounded with respect to their System 1-type intuitions, especially if the intuition is affectively charged.

It is also possible, in some cases, for judgments arrived at initially through System 2-type reasoning to become "automatized" in System 1-type processes (Bargh & Chartrand, 1999). This is common with expert judgments. In these cases, a person makes consciously reasoned System 2-type judgments that, with practice and experience, leads to know-how, which can produce judgments without having to go through any conscious steps of reasoning. In these cases, an intuitive judgment can be based in reasons, though the person may not be consciously aware of the reasons for that judgment. As I said before, it is possible that some intuitive moral judgments are

the products of such a process of automatization, but it is not plausible to think that all intuitive moral judgments are the products of such a process.

This dual-process architecture of human reasoning and judgment, with two different and causally distinct types of processes that interact in some important ways, provides the right sort of general framework for moving beyond the Unity of Process Thesis with respect to moral judging and judgment. If it is generally the case the humans can arrive at a judgment in some domain through two distinct types of cognitive processes, then it is quite possible that moral judgments can be arrived at through two distinct types of cognitive process, both of which are capable of producing genuine moral judgments through their own distinctive types of operations, and which can also interact in some complex ways. I shall explore this possibility in the next section, but before I do one caveat is necessary. New wave sentimentalists often present their models of moral judgment as "dual-process" models, because they give a role to both the emotions and reasoning. As a result, one might think that such accounts are *eo ipso* not constrained by the Unity of Process Thesis. But this is not the case. The Unity of Process Thesis is not a claim about *how many* processes can causally contribute to the production of some moral judgments. There could be one, or two, or a thousand. It makes no difference. The claim of the Unity of Process Thesis is that only one of those processes is the "core" psychological process that can produce genuine moral judgments. While both Social Intuitionists and Constructive Sentimentalists give both the emotions and reasoning a role in moral judging, they both maintain that the "core" psychological process of moral judging is the emotions, and that reasoning serves only as an aid to the emotions; either by triggering new

emotional responses through post hoc reasoning, or by categorizing actions or persons. Neither Social Intuitionism nor Constructive Sentimentalism allow that there are two distinct types of cognitive processes that are each capable of producing a genuine moral judgment through their respective and distinctive operations.[11] Consequently, neither is consistent with the general architectural claims of dual process theory. Hence, even though they claim to be dual process theories of moral judgment, they are not.

5. A Dual-Process Architecture for Moral Judgment: The Two Kinds Hypothesis

Starting from the general outlines of dual-process architecture of human judgment, it is possible to sketch a model of moral judging that moves beyond the Unity of Process Thesis. Recall that the Unity of Process Thesis is the claim that all genuine moral judgments are the products of a single "core" psychological process. Under the constraints of the Unity of Process Thesis, the theoretical space for thinking about the underlying processes of moral judging is a choice between emotions or reasoning. However, the dual-process architecture of human judgment suggest that it is possible that moral judging is better explained by a dual-process architecture; one that takes it that System 1-type processes and System 2-type processes, operating independently, can produce genuine moral judgments, and that the empirical research that points to a large causal role for the emotions in moral

[11] Social Intuitionists may claim that this characterization of their model is unfair, because they allow that moral judgments can be arrived at through "the sheer force of logic, overriding [an] initial intuition" (Haidt and Bjorklund, 2008a, pg. 193). However, in the very next paragraph they argue that such judgments have no actual role in moral thinking or acting, which is another way of saying that they are not genuine moral judgments; they are much more like Prinz's notion of a verbalized moral judgment. They are words that could be used to express a moral judgment (and words a person may mentally assent to), but they are not genuine moral judgments.

judging might illuminate the underlying operations of one of these two types of processes and not the other. Moreover, by placing moral judging and judgment within the broader context of the dual-process architecture of human judgment in general there is no reason to think that only one or the other of these two processes is the single "core" process of moral judging that alone can produce genuine moral judgments. Under such a dual-process model of moral judging there would be no single "core" psychological process of moral judging; there are two, and any adequate model of moral judging will need to provide an account of both of these distinct processes. I shall attempt to provide a sketch for how this can be done, but it is important to recognize that what I am proposing here is more of a framework for future research and not a fully developed model of moral judging. It is an attempt to clarify what is often conflated by those working under the constraints of the Unity of Process Thesis, and show what an actual dual-process model of moral judging might look like. I call the framework I propose the Two Kinds Hypothesis.

The first step in developing the Two Kinds Hypothesis is to show that various features moral judging and judgment require an explanation in terms of a dual-process architecture. Recall that proponents of dual-process theories argue that dual-process theories are necessary for explaining the differences between the processes that lead to quick intuitive judgments and the processes that lead to reasoned judgments, which can be labeled intuitive and deliberative judgments, respectively. One question, then, is whether there is some phenomenon with respect to moral judging that requires a similar explanation in terms of two distinct types of processes. And indeed, many philosophers already distinguish between quick intuitive moral

judgments (or moral intuitions) and more reasoned deliberative moral judgments. However, to determine whether this distinction between intuitive and deliberative moral judgments requires an explanation in terms of a dual-process architecture, it is important to flesh out more precisely what the properties of intuitive and deliberative moral judgments are supposed to be, and whether they correspond with the general claims of dual-process theory.

Starting with intuitive moral judgments, Mikhail (2010) characterizes intuitive moral judgments as spontaneous, stable, involuntary, and immediate moral judgments. Sinnott-Armstrong et al. characterize moral intuitions as "strong, stable, and immediate moral beliefs" (Sinnott-Armstrong, Young, & Cushman, 2010, p. 1), and Haidt and Bjorklund (2008a) define moral intuitions as "the sudden appearance in consciousness, or at the fringe of consciousness, of an evaluative feeling (like-dislike, good-bad) about the character or actions of a person without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion" (pg. 188). Though there are some divergences in these characterizations of intuitive moral judgments, there is also some striking overlap. Intuitive moral judgments are arrived at quickly, without any intervening steps of conscious thinking, weighing of reasons, application of a moral theory, or any such thing. In the first instance, it is these characteristics of intuitive moral judging that require an explanation in terms of the underlying psychology of intuitive moral judging.

Moreover, these characteristics reveal that intuitive moral judging is an *automatic* process, meaning that is totally involuntary—it is simply something that minds like ours do regardless of any conscious decision to do so, in much the same

way that minds like ours automatically "see" objects in our visual field regardless any conscious decision or effort to do so. Indeed, intuitive moral judgments are often described using visual metaphors such as "moral perception" or simply "seeing as" (Aristotle, 1999; Blum, 1991), in an attempt to capture something of the automaticity of intuitive moral judging. The processes of intuitive moral judging are also psychologically *immediate*, meaning that they do not involve any conscious steps of reasoning or inference. Intuitive moral judgments just "land" in consciousness with a perception-like quality. Again, visual metaphors are useful. People simply "see" certain situations, actions, or people as good or bad, right or wrong, without any conscious effort (Harman, 1997). This is precisely what happens when people are presented the Trolley and Footbridge cases; they often make an immediate moral judgment without any conscious effort.[12]

Moreover, because intuitive moral judgments land in consciousness without any intervening conscious steps of reasoning, people can be dumbfounded with respect to these judgments when pressed to provide reasons for them. Because intuitive moral judgments are not directly based in conscious reasoning, the reasons people provide for them can *only be* post hoc justifications, some of which can be irrelevant or insufficient to actually support the judgment.[13] Even in situations where people's reasons are irrelevant or insufficient, and even after this has been pointed out

[12] Greene et al. argue that some people are able to slow down and make a more calculated utilitarian judgment when presented the Footbridge case (Greene, et al., 2001). This is based their finding that the mean response time for those who judge that it is "appropriate" to push the fat man in front of the trolley in the Footbridge case is longer than the mean response time for those who judge that the action is "inappropriate." McGuire et al, however, criticize this interpretation of the mean response time data (McGuire, Langdon, Coltheart, & Mackenzie, 2009). They argue that using individual comparisons (as opposed to between group comparisons) eliminates the statistical significance.

[13] Post hoc justifications are not always irrelevant or insufficient. Some post hoc reason actually can provide appropriate rational support for an intuitive moral judgment.

to them, they may continue to maintain their intuitive moral judgments in just the same way people maintain their intuitive rank orderings of Linda's likely career paths even when it is pointed out to them that such rank orderings violate the rules of conjunction. What moral dumbfounding makes quite clear is that the processes of intuitive moral judging (and intuitive judging in general) are *nonconscious* and *opaque*, meaning that their internal workings are not conscious, nor are they introspectively accessible.

Deliberative moral judgments, on the other hand, are the products of some episode of conscious reasoning; weighing reasons, searching for evidence, and inferring a conclusion. Deliberative moral judging is a *conscious* process, involving conscious steps of reasoning and inferring a conclusion. As such, the process is not opaque,[14] and the conclusions are certainly not immediate. Deliberating over a moral question can be a hard slog and quite *time-consuming*, taking hours, days, or longer involving the consideration of arguments, counter-arguments and a myriad of other possible concerns. Moreover, the processes of deliberative moral judging are under *voluntary* control. A person can decide when and if to deliberate about some moral question, though whether a person engages in such deliberation will be determined, in part, by his or her deliberative temperament, time constraints, and other exogenous factors. Importantly, deliberation is not something that minds like ours simply do—it is something a person must decide to do.

[14] There are two important caveats here. Sometimes drawing a connection between two ideas, or cases, or reasons can just "pop" into one's mind. For example, a person may be deliberating on a moral question when a particular solution or conclusion or line of argument simply bubbles up to consciousness. The processes underlying such eureka moments are certainly opaque, and likely involve some processes of intuitive moral judging. Second, while System 2-type deliberation involves conscious steps of reasoning, not all System 2-type processing is conscious or consciously accessible.

Moreover, because the processes of deliberative moral judging are conscious, and involve consciously considering reasons and arguments, deliberative moral judgments are not "dumbfoundable" in the same way that intuitive moral judgments are. The reasons and arguments that people give for deliberatively judging a case a certain way can and should be the reasons and arguments that led them to that particular moral conclusion.[15] And if people's moral arguments are sound, then those arguments are rationally compelling on others as well. A person can expect others to draw the same deliberative conclusion, and deliberatively judge the case the same way she has, or show where her reasoning went wrong.

Intuitive and deliberative moral judging, then, can be distinguished in terms of a set of contrasting properties: deliberative moral judging is voluntary, time-consuming, conscious, and based in consciously accessible reasons. Intuitive moral judging, on the other hand, is automatic, quick, nonconscious, and opaque. Table 5.2 summarizes these differences.

[15] If the reasoning involved in deliberatively judging a case one way is complicated, it is quite possible that a person might not be able to cite all the reasons and arguments that support deliberatively judging a case a certain way, but a person should still be able to offer at least some of the reasons that they actually considered in drawing the moral conclusion that they did.

Deliberative Moral Judging	Intuitive Moral Judging
Voluntary	Automatic
Conscious	Nonconscious
Time-consuming	Quick
Introspectively accessible reasons	Opaque

Table 5.1: Summarizes the differences between intuitive and deliberative processes in moral judging.

Remember that the distinction between intuitive moral judgments and deliberative moral judgments is widely acknowledged among moral philosophers, and the differences between intuitive moral judging and deliberative moral judging appear to require an explanation in terms of different underlying psychological processes. Moreover, the differences between intuitive and deliberative moral judging fit very nicely with the general outlines of the dual-process architecture of human judgment. Just like intuitive moral judging, System 1-type processes are quick, automatic, nonconscious, opaque, and dumbfoundable. Similarly, deliberative moral judging matches the features of System 2-type processes of being time-consuming, voluntary, conscious, and involve introspectively accessible reasons. This strongly suggests that a dual-process architecture of moral judgment is the best explanation of the distinction between intuitive and deliberative moral judging —a suggestion which gains further support when one considers that attempting to explain the differences between intuitive and deliberative moral judging by a unified theory of moral judging

under the constraints of the Unity of Process Thesis raises serious theoretical and explanatory worries. A dual process theory of moral judging can avoid these worries because it does not require that all episodes of moral judging be explained in the same way. There is no forced dichotomous choice between reasoning and emotions. A dual-process architecture of moral judging opens the possibility that some episodes of moral judging are explained, for example, by the emotions, while other episodes of moral judging are explained, for example, by reasoning. I shall return to this possibility shortly, but first I want to give two more reasons for thinking that a dual-process architecture provides the right framework for understanding the distinction between intuitive and deliberative moral judging.

First, a dual-process architecture of moral judgment can straightforwardly explain the potential conflict between moral judgments and the requirements of a moral theory or moral principles that one would consciously endorse. One point that emerged in the previous discussion of reflective equilibrium is that moral judgments can conflict with requirements of a moral theory or moral principles that one would consciously endorse, or can fail to cohere in the right sort of way with one's more reasoned moral judgments. The most straightforward explanation of this potential conflict is that many of the moral judgments that conflict with the requirements of a moral theory or moral principles that one would consciously endorse are intuitive moral judgments that are the products of System 1-type processes, which are quick, automatic, opaque, and nonconscious. Thus, the conflict between intuitive moral judgments and the requirements of a moral theory or moral principles that one would consciously endorse can be explained in just the same way as the conflict between

people's intuitive judgments in the Linda case and the normative requirements of probability theory that they consciously endorse: they are the products of two distinct types of cognitive process. Only with such a dual-process architecture of moral judging is it possible to explain how it is people reflectively distance themselves from their moral judgments to ask whether they cohere in the right way with their consciously held moral principles and other non-moral beliefs.

Second, a dual-process architecture of moral judging can explain how it is that conscious reasoning can, and can fail to, override an intuitive moral judgment that conflicts with the requirements of a moral theory or moral principles that a person consciously endorses. According to dual-process theory, in some cases it is possible for System 2-type reasoning to override a System 1-type intuition, but such overriding is difficult and sometimes fails. A similar situation holds in the moral case as well. People sometimes recognize that their initial moral judgment conflicts with, or is unsupported by, their consciously endorsed moral theories or moral principles, and are able to override their initial intuitive moral judgment after having reasoned through it. Sometimes, through habit, this override becomes reflexive, and the person will typically think and act consistently with their more reflective moral judgment.

However, there are cases when people are unable to override an initial intuitive moral judgment, even though they recognize good reasons to do so. A dual-process architecture of moral judging provides a straightforward explanation of how this is possible, too. System 2-type reasoning does not always succeed in overriding an initial judgment, in part because doing so requires a great deal of cognitive resources that are not always available, and also in part because some people are

more likely to trust their "gut instinct" more than their more carefully reasoned judgments when the two conflict. In some cases, when people cannot override an intuitive moral judgment, they may have a dual-attitude, where they have a conflicting intuitive and deliberative moral judgment. In such cases, when people act on a deliberative moral judgment that conflicts with an intuitive moral judgment, they may feel regret, even though they recognize that they have done the best they could, or what was morally required, all things considered. This is sometimes referred to as "moral residue," and indicates that some intuitive moral judgments cannot be entirely overridden by conscious reasoning, and can have lingering effects even when a person recognizes good reasons to act differently (Williams, 1973b, p. 176).

The distinction between intuitive and deliberative moral judging and judgments, therefore, maps nicely onto the general claims of the dual-process architecture of human judgment. Moreover, a dual-process architecture of moral judging provides a very straightforward framework for explaining the common distinction between intuitive and deliberative moral judging and judgments, and it avoids the theoretical and explanatory problems that follow from the Unity of Process Thesis. A genuine dual-process framework for moral judging, such as the Two Kinds Hypothesis, is therefore worth serious consideration. And importantly, then, the task of providing a model of moral judging requires providing a model for both the System 1-type processes of intuitive moral judging and the System 2-type processes of deliberative moral judging, and giving some account for how these processes interact.

There is one very important consequence to this last claim, which is that empirical findings with respect to moral judging must be indexed to the type of moral judging being studied. That is, empirical findings cannot properly be interpreted as providing insight into moral judgment, *simpliciter*; rather they provide insight into either intuitive moral judging or deliberative moral judging, or the interactions between these two types of process. It is important to bear this in mind, because most of the empirical findings cited by Social Intuitionists and Constructive Sentimentalists in support of their models of moral judging apply most directly to the System 1-type processes of intuitive moral judging. The research paradigms employed in behavioral studies and brain scans are meant to elicit a quick moral judgment in response to a vignette or scene. Subjects are not given time to reason through the vignettes or scenes carefully, but are asked to answer within very narrow timeframes before they move on to the next task (sometimes as short as 2-3 seconds). Importantly, then, such research elicits people's quick, intuitive moral judgments, and so such paradigms mostly illuminate only the inner-workings of System 1-type intuitive moral judging. Therefore, the correct interpretation of these findings, *contra* Social Intuitionists and Constructive Sentimentalists, is not that moral judgment *simpliciter* is easily influenced by the emotions, and that emotion centers of the brain light up when people make moral judgments, but that intuitive moral judging is easily influenced by the emotions and that the processes of intuitive moral judging are correlated with certain emotion centers in the brain.

Because Social Intuitionists and Constructive Sentimentalists fail to appreciate the need to distinguish between two distinct types of process in moral judging when

interpreting empirical findings, they tend to overstate the findings from behavioral studies and brain scans and other, similar research paradigms by arguing that the reveal something about the processes of moral judgment *simpliciter*. Understood in this broad way, this research would tend to support their claims that moral judging is largely a matter of emotions as opposed to reasoning. However, the same research, understood in a more limited way, when indexed to intuitive moral judging, only reveals that emotions have some causal role in System 1-type intuitive moral judging.

Moreover, given the limitations of the extant empirical literature, it is hard to say with any confidence what the precise causal role of the emotions in System 1-type intuitive moral judging is. It could be that emotions are sufficient to cause an intuitive moral judgment (Greene, 2007; Greene & Haidt, 2002; Haidt & Bjorklund, 2008a), or that they partly constitute an intuitive moral judgment (Prinz, 2007), or that the processes of intuitive moral judging cause an emotion consistent with the moral judgment (Hauser, 2006; Sripada & Stich, 2006), or that emotions help focus pre-conscious attention by rendering certain features of a scene particularly salient (Craigie, 2011; Huebner, et al., 2009).[16] All of these causal stories are consistent with the extant empirical literature, but regardless of these differences there is a general consensus that the emotions are causally involved in the quick, automatic, opaque and nonconscious processes of intuitive moral judging, and that intuitive moral judgments are, to some extent, affective judgments. It could be that the processes of intuitive moral judgment token moral emotions, such as anger, shame, or guilt apart from the

[16] Some of these positions may fare better with respect to the developmental literature than others. There are a number of studies that suggest intuitive moral judging develops early and with a predictable ontology, which favors a nativist explanation of intuitive moral judging, such as a Universal Moral Grammar or Social Intuitionism, as opposed to empiricist views such as Constructive Sentimentalism.

judgment itself, or that intuitive moral judgments are partly constituted by the emotions, with the emotions being bound to the judgment. This latter possibility is more likely, because emotions are typically bound to some sort of representation, that is, they typically take objects towards which the emotion is directed (de Sousa, 2003; Greenspan, 1988; Oakley, 1992). In most cases, people are not just angry, for example, they are angry at something for some reason, which is the particular object of their anger. Moreover, this latter possibility helps explain one sense of the notion of a "thick" moral judgment, where the cognitive and affective aspects of the moral judgment cannot be pulled apart (Williams, 1986; Zagzebski, 2003). However, more fine-grained empirical research is needed to make more definitive claims with respect to the causal role of the emotions in intuitive moral judging.

Regardless, the important point is that the emotions play some important causal role in intuitive moral judging, but it is unlikely that the emotions play the same causal role with respect to System 2-type deliberative moral judging. System 2-type deliberative moral judging generally involves weighing and considering reasons and arguments, including consciously endorsed moral theories or moral principles, non-moral beliefs, and even one's System 1-type intuitive moral judgments, to determine whether they cohere in the right sort of way, and which of one's moral commitments, attitudes, and judgments are justified, all things considered. In these ways deliberative moral judging is no different from conscious reasoning in other domains, which is why it is common for moral philosophers to claim that "Moral reasoning is ordinary critical reasoning applied to ethics" (Vaughn, 2008, p. 43). Moreover, just as other System 2-type processes, deliberative moral judging is

conscious, time-consuming, effortful, and subject to ordinary norms of rationality. It is also difficult to do well, though it improves with general improvement in reasoning skills (Colby, Kohlberg, Gibbs, & Lieberman, 1983; Lapsley, 1996).

Moreover, unlike System 1-type intuitive moral judging, System 2-type deliberative moral judging is not automatic, and so whether a person actually undertakes System 2-type moral deliberation to produce an all-things-considered deliberative moral judgment is going to be determined, in part, by time pressures, cognitive load, being tired, annoyed, or distracted, and similar endogenous and exogenous factors. But again, the important point is that System 2-type deliberative moral judging involves domain general reasoning, and consequently deliberative moral judgments are not emotional judgments in the same way that System 1-type intuitive moral judgments are. They are more like ordinary beliefs, and are, in one sense, "thin" moral judgments in that they are cognitive judgments that lack an affective component (Williams, 1986; Zagzebski, 2003).

Therefore, according to the Two Kinds Hypothesis there is a role for both emotions and reasoning: emotions play an important causal role in System 1-type intuitive moral judging, but they do not play the same role in System 2-type deliberative moral judging. Moreover, System 2-type deliberative moral judging typically involves domain general reasoning to reach a moral conclusion, whereas System 1-type intuitive moral judging does not, and could even rely on processes specific to the moral domain (Cosmides, 1989; Dwyer, 2006, 2009; Hauser, 2006; Hauser, Young, & Cushman, 2008; Mikhail, 2007, 2009). This clean division between the two processes, however, is something of an abstraction, and belies the

complex interactions between the System 1-type intuitive moral judging and System 2-type deliberative moral judging. For example, System 1-type intuitive moral judgments often serve as the starting points for System 2-type deliberative moral judging. Most people rarely undertake System 2-type deliberative moral judging with respect to some question in the absence of a System 1-type intuitive moral judgment, except perhaps in the context of a philosophy class, just as Social Intuitionists maintain. Moreover, System 2-type deliberative moral judging is often "anchored" by one's intuitive moral judgments. That is, not only do intuitive moral judgments generally serve as a starting point for deliberation, but one's intuitive moral judgments can drive deliberation in a certain direction, for example, by biasing a person towards confirming evidence and away from disconfirming evidence (Baron, 2003).

On the other hand, System 2-type deliberation can influence one's System 1-type intuitive moral judging, by drawing attention to factors in a situation that elicit a different System 1-type intuitive moral judgment, just as Social Intuitionists and Constructive Sentimentalists claim. Importantly, however, System 2-type deliberative moral judging cannot directly influence the internal workings of the System 1-type processes of intuitive moral judging. Moreover, in some cases System 2-type deliberation can override a System 1-type intuitive moral judgment, though doing so requires a great deal of cognitive energy, and even then is not always successful. As I said before, a person can experience "moral residue" when acting on a System 2-type all-things-considered deliberative moral judgment that conflicts with a System 1-type intuitive moral judgment, which indicates that although the person successfully acted

on a System 2-type deliberative moral judgment, the System 1-type intuitive moral judgment persisted and became a source of guilt. Other times, a System 1-type intuitive moral judgment can be successfully overridden by System 2-type deliberative moral judgment, perhaps even automatically, and a person will think and act consistently with their System 2-type deliberative judgment without residue, though the System 1-type intuitive moral judgment can resurface in cases where System 2-type reasoning is blunted or disrupted, such as after a hard night of drinking, or being tired or distracted.

Therefore, even though it is possible to distinguish between System 1-type intuitive moral judging, and System 2-type deliberative moral judging, in many cases a particular episode of moral thinking will involve both types of processes, interacting in some complex ways. Determining more precisely how these interactions are realized will require more empirical research, but the interactions between System 1-type intuitive moral judging and System 2-type deliberative moral judging posited by the Two Kinds Hypothesis are consistent with the general claims of dual-process architecture of human judgment. Moreover, such interactions help explain broader aspects of our moral psychology, including moral change and moral justification. As I argued before, moral change and moral justification both involve a back-and-forth between moral intuitions and moral reasoning, and a dual-process architecture of moral judging helps to explain how such interactions are possible.

6. From Dual Processes to Two Kinds

A dual-process architecture of moral judging provides the right framework for explaining the differences between intuitive and deliberative moral judging and

judgment, and the complex interactions between the two helps explain broader aspects of our moral psychology. System 1-type intuitive moral judging is quick, automatic, nonconscious, opaque, and involves the emotions in important ways. System 2-type deliberative moral judgment is slow, effortful, conscious, voluntary, and does not involve emotions in the same way. Not only do emotions figure differently in the two systems, System 1-type intuitive moral judging and System 2-type deliberative moral judging play different roles with respect to moral justification and moral change. This suggests the possibility that intuitive and deliberative moral judgments function differently in our moral psychology with respect to thinking and acting, and thus that they are not simply products of two different types of process, but that they are distinct psychological kinds as well.

The most straightforward support for this claim is the different phenomenological qualities of intuitive and deliberative moral judgments. Ordinary intuitive moral judgments have a certain motivational "oomph," a "to-be-doneness" about them that ordinarily moves people to act consistently with them (Mackie, 1977). In general, intuitive moral judgments *feel* different than merely descriptive judgments: they compel us, they move us, they have a certain directedness towards action. Moreover, they are not "flat" or "cool" judgments: they are often quite "hot." This is metaphorical language, but it captures something of the feeling of intuitive moral judgments. Moreover, part of the phenomenology of intuitive moral judgments is that they feel universal and categorical; meaning that they apply to everyone and that their demands cannot be escaped or laid aside for trivial reasons—or for no reason at all. To make an intuitive moral judgment is to feel obligated by that moral

judgment, and to feel guilt or shame if one fails to satisfy that obligation (Strawson, 1962), which is precisely why a person can feel "moral residue" even acting an all-things-considered deliberative judgment that conflicts with an intuitive moral judgment.

The phenomenological feel of an intuitive moral judgment is part of our ordinary moral experience, and it indicates that there is an affective component to ordinary intuitive moral judgments because emotions have a similar phenomenological feel. Emotions feel a certain way, have a certain sort of "oomph" and directedness towards action, and can often be quite "hot." This is precisely why sentimentalists often stress the similarities between emotions and moral judgments (D'Arms & Jacobson, 2000b; Stevenson, 1937). Stevenson, for example, argues that the moral use of the term "good" has a particular emotive meaning, which is the "tendency of a word, arising through the history of its usage, to produce (result from) *affective* responses in people. It is the immediate aura of feeling which hovers around about a word" (Stevenson, 1937, reprinted in Darwall, Gibbard, & Railton, 1997, p. 77). And phenomenologically, intuitive moral judgments do have this "aura of feeling" about them.

Deliberative moral judgments, on the other hand, do not; at least not in the same way. It is not at all uncommon for ordinary people to come to some moral conclusion through deliberation and not feel directly moved by it. Perhaps they have been puzzling through some complicated issue by applying moral principles, weighing various options, and coming to an all-things-considered moral judgment. Such moral judgments are not "hot" in the same way that intuitive moral judgments

are, and they do not have the strong pull of "to-be-doneness" or the "oomph" that characterizes intuitive moral judgments. Many people, for example, have been convinced by Singer's argument that it is morally wrong to eat farm-grown meat (Singer, 1974), and deliberatively judge that it is wrong to eat meat, but this moral judgment has no direct motivational or affective effects. It could be argued that such people do not make a genuine moral judgment, but for reasons given before, this is not the best interpretation. Rather, the better interpretation is that deliberative moral judgments, arrived at through System 2-type reasoning alone, do not come with the "oomph" or feel that are normally associated with intuitive moral judgments. Indeed, it is precisely this variation in the motivational effects among a person's moral judgments that undermines the plausibility moral judgment internalism (Svavarsdottir, 1999).

So, intuitive and deliberative moral judgments have different phenomenological feels, different motivational effects, and as I argued in Chapter 4, play different roles with respect to moral justification and in moral thinking. Thus, intuitive and deliberative moral judgments play different roles in our moral psychology, and thus they are justifiably considered distinct psychological kinds. This is why my view is called the Two Kinds Hypothesis.

At this point one might object that the differences between intuitive and deliberative moral judgments is more of a reason to conclude that only one kind of moral judgment, either intuitive or deliberative, can be a genuine moral judgment. But this move is hard to square with the simple methodological principle that robust phenomenological features of judgments are better explained as the product of a

capacity, rather than some deep-seated cognitive error that would render such judgments not genuine (Horgan & Timmons, 2007).[17] It should be noted that this methodological claim is entirely consistent with the finding that some capacities are subject to well-known performance errors that give rise to predictably erroneous judgments, such as in the Linda case. The phenomenology of a judgment is quite distinct from the content of that judgment. The methodological principle I am endorsing fixes solely on the phenomenology, and intuitive and deliberative moral judgments have strikingly different phenomenologies. Thus, the better explanation of the different phenomenological qualities, motivational effects, and roles in a person's moral psychology between intuitive and deliberative moral judgments is that they are both the products of proper functioning capacities for moral judging, and thus that they are both genuine moral judgments, rather than the competing claim that only one kind of moral judgment is genuine. And, if that is the case, then intuitive and deliberative moral judgments are best thought of distinct psychological kinds. Moreover, the principal reason to reject the conclusion that intuitive and deliberative moral judgments are distinct psychological kinds is to defend metaethical, as opposed to psychological, claims (see, for example, Prinz, 2007; Smith, 1994).

[17] Horgan and Timmons call this methodological principle the *maxim of default competence-based explanations*. According to this maxim:

> All else equal, a theoretical explanation of a pervasive, population-wide, psychological phenomenon will be more adequate to the extent that (1) it explains the phenomenon as the product of cognitive competence rather than as a performance error, and (2) it avoids ascribing some deep-seated, population-wide, error-tendency to the cognitive architecture that subserves competence itself (e.g., an architecturally grounded tendency to erroneously conflate post-hoc confabulation with articulation of the actual reasons behind one's moral judgments) (Horgan & Timmons, 2007, p. 289).

It is important to note, however, that the distinction between intuitive and deliberative moral judgments as distinct psychological kinds is not pure, in the sense that it would be possible for any moral judgment to trace the causes of that moral judgment solely to the operations of either intuitive or deliberative moral judging. Moral judging is not nearly so neat and tidy, and many episodes of moral judging will involve the back-and-forth of System 1-type intuitive processes and System 2-type deliberative processes. However, it would be wrong to conclude from this caveat that the distinction between intuitive and deliberative moral judgments is simply a fiction; rather, it is a useful idealization that provides a necessary framework for capturing the different roles intuitive and deliberative moral judgments play in a person's moral psychology. In the end it might not make sense to ask whether a particular judgment is an intuitive moral judgment or a deliberative moral judgment, but rather how intuitive it is versus how deliberative it is. Pure intuitive moral judgments set one end of the spectrum, while pure deliberative moral judgments set the other end of the spectrum, though most moral judgments will fall somewhere in between. This is precisely why many ordinary deliberative moral judgments are not entirely affectively flat, because they likely involve some System 1-type intuitive processing.

Again, the Two Kinds Hypothesis is meant as a framework for the study of moral judging and judgment. It is not a fully developed model of moral judging, and it is not intended as such. Rather, it lays out the general contours of what an adequate model of moral judging should look like by drawing on a consensus view of the dual-process nature of human judgment in general. Continuing empirical research is needed to fill in this general framework, and the Two Kinds Hypothesis is useful in

providing a framework for directing such research. Not only does the Two Kinds Hypothesis provide a useful framework for empirical research, it also provides an incredibly useful framework for finding a way forward in some seemingly intractable debates in metaethics, including moral motivation, and the debate between moral particularists and generalists. I shall turn to these metaethical issues in the next chapter, and show how the Two Kinds Hypothesis provides a way forward, and, moreover, that these metaethical problems emerge only if one accepts a closely related Thesis to the Unity of Process Thesis, the Unity of Kind Thesis.

7. Conclusions

The primary motivation for the Two Kinds Hypothesis is the explanatory inadequacy of current models of moral judging that implicitly accept the truth of the Unity of Process Thesis. Under the constraints of the Unity of Process Thesis, the underlying process of moral judging is seen as the choice between reason and the emotions. Dual-process architecture of human judgment provides the right general framework for moving beyond the Unity of Process Thesis with respect to moral judging, and provides a framework for distinguishing between System 1-type processes of intuitive moral judging, and System 2-type processes of deliberative moral judging. The Two Kinds Hypothesis provides such a framework, though filling it in will require additional empirical and theoretical work. One consequence of the Two Kinds Hypothesis framework of moral judging is that intuitive and deliberative moral judgments should be distinguished as distinct psychological kinds. This has important implications for some debates in metaethics, which I shall explore in the next chapter.

Chapter 6: Implications for Metaethics

Part of the interest in models of moral judging is that they often have implications for debates in metaethics, which focus on a range of metaphysical, epistemological, semantic, and psychological questions with respect to moral judgments and morality. However, debates in metaethics tend to be organized and conceptualized around universal claims with respect to moral judgments, and underlying this way of organizing and conceptualizing these debates is the assumption that all genuine moral judgments are the same psychological kind. Call this assumption—that moral judgments are a single psychological kind—the Unity of Kind Thesis. Importantly, the Unity of Kind Thesis and the Unity of Process Thesis are mutually supporting. If moral judgments are products of a single "core" psychological process, it is natural to assume that this process produces a single kind of moral judgment that is a token of the same kind of mental state and plays the same role in the mental economy. Similarly, if one takes it that moral judgments are a single psychological kind, then one only needs to find a single "core" psychological process capable of producing them. Thus, the two theses are mutually supporting, though strictly speaking, neither entails the other.

The Two Kinds Hypothesis challenges both the Unity of Process Thesis and the Unity of Kind Thesis. And because debates in metaethics tend to be organized and conceptualized around the Unity of Kind Thesis, if the Two Kinds Hypothesis is correct it opens up an interesting theoretical possibility that debates in metaethics have been organized and conceptualized on the wrong terms. Metaethicists have

assumed of these debates that each depends upon a choice between two exclusive possibilities, for example, that moral judgments either motivate directly, or indirectly; or that moral thinking either always involves general principles, or it never does. But if the Two Kinds Hypothesis is correct, these seemingly contrary positions might not represent actual contraries, because it might be possible that one claim is true of one kind of moral judgment, while another claim is true of another kind of moral judgment. Universal claims with respect to moral judging and judgments need to be indexed to the kind of moral judgment under discussion, either intuitive or deliberative, and doing so can provide a valuable way forward in some metaethical debates. That, at least, is what I intend to show. But given the complexities of metaethical debates, and given the ways they have been organized and conceptualized, this is best done by looking at only a small subset of these debates to show how the Two Kinds Hypothesis can helpfully reorganize and reconceptualize the terms of these debates to provide a new way forward. I shall focus here on two debates in particular: the debate between internalists and externalists with respect to moral motivation; and the debate between moral particularists and moral generalists with respect to the role of moral principles in moral thinking.

1. Moral Motivation

One thing that distinguishes moral judgments from other sorts of judgments is that they are reliably connected to motivation. More precisely, when people make a sincere moral judgment they are reliably motivated to act consistently with it, at least in ordinary people in the ordinary case. Even though there is a reliable connection between a moral judgment and the motivation to act consistently with it, this

motivation is by no means always overriding, nor does it guarantee that a person will actually act consistently with their moral judgments.[1] One central topic in metaethics is explaining how moral judgments are connected to motivation in this way, and just as importantly, how it breaks down. In broad strokes, there are two possible ways to account for the reliable connection between a moral judgment and motivation: *moral judgment internalism* and *moral judgment externalism* (henceforward internalism and externalism, respectively). Internalists claim that the connection between moral judgments and motivation is internal to the moral judgment itself, either because moral judgments are partly constituted by motivational states (Gibbard, 1990; Prinz, 2007; Williams, 1981), or because the proper application of moral concepts, such as good, right, permissible, impermissible, etc, necessarily involves a motive to act because it is part of what these concepts *mean* (Dreier, 1990; Korsgaard, 1996; Smith, 1994). Externalists, on the other hand, claim that the connection between a moral judgment and motivation is neither necessary nor conceptual, but is contingent on the moral judger having an appropriate desire or conative attitude under a moral mode of presentation, such as the desire to be moral, or the desire to do the right thing, or some cluster of desires or attitudes that provides the motivation to act consistently with a moral judgment.[2]

[1] Some philosophers argue that moral judgments always provide an overriding reason, and so someone who fails to act consistently with a moral judgment is necessarily irrational (e.g., Smith, 1994). This is a substantive claim about the relationship between morality, reasons, and rationality, not a psychological claim about whether moral judgments always override a person's other motives. Since it is the psychology under investigation here, this sort of claim can be set aside. It does not matter here whether such a person is rational or irrational, only that such failure of moral motivation is psychologically possible.

[2] It is important to distinguish moral judgment internalism and moral judgment externalism from other positions in metaethics that are sometimes called internalism and externalism. Reasons internalism is the view that reasons necessarily motivate, while reasons externalism is the view that reasons do not necessarily motivate. Recall that many moral philosophers view moral judgments as judgments related

In attempting to explain the reliable connection between a moral judgment and the motivation to act consistently with it, both internalists and externalists implicitly assume the truth of the Unity of Kind Thesis—the claim that all moral judgments are a single psychological kind—and thus they both assume that there is a single explanation for this reliable connection that holds for all moral judgments. However, it is simply not possible to explain the reliable motivational force of all moral judgments in the same way without significant theoretical and explanatory losses. By rejecting the Unity of Kind Thesis, it is possible to move to forward in this debate, which can be seen by investigating one very influential internalist argument.

Smith provides one very influential internalist argument is meant to show that there is a straightforward *reductio* of the externalist view of moral motivation in the ordinary case (Smith, 1994). Smith here pursues a strategy of indirect proof: if externalism can be shown to be false (via *reductio*), then internalism must be true. To get the *reductio* started, Smith presents the following example of ordinary moral motivation:

> Suppose I am engaged in an argument with you about a fundamental moral question; a question about whether we should vote for the libertarian party at some election as opposed to the social democrats. In order to make matters vivid, we will suppose that I come to the

---

in the right way to what reasons one has to act a certain way or hold an attitude of a certain kind. When it comes to accounting for the various motivational effects and failures of moral judgments, the challenge for philosophers who adopt this view is providing such an account purely in terms of the structure of practical reasons. This debate centers around the question of what sorts of mental states can motivate (beliefs, desires, or both), and whether reasons necessarily involve desires (Shiffrin, 1999), give rise to desires (Smith, 1994; Williams, 1981), or are motivating beliefs (Nagel, 1979). But the questions of whether moral judgments give rise to a reason to act a certain way or hold an attitude of a certain kind can be separated from the question of how moral judgments motivate (McDowell, 1988a). The first question involves the relationship between moral judgments and reasons, while the second concerns the relationship between moral judgments and motivation.

> argument already judging that we should vote for the libertarians, and already motivated to do so as well. During the course of the argument, let's suppose you convince me that I am fundamentally wrong. I should vote for the social democrats, and not just because the social democrats will better promote the values I thought would be promoted by libertarians, but rather because the values I thought should and would be promoted by libertarians are fundamentally mistaken. You get me to change my fundamental values…If I am a good and strong-willed person then a new motivation will follow in the wake of my new judgment (1994, pp. 71-72).

What is the best explanation of this change in motivation? Smith argues that there are two possibilities: either the motivation follows directly from the conceptual content of the moral judgment (the internalist explanation); or it follows from the person's other already existing motives and dispositions (the externalist explanation). If one takes the externalist line, then the motivation is mediated by some desire or other conative state, such as the desire to be moral. Which means that for the externalist, people do not have a direct concern for the objects of their moral judgments—they do not care directly about being kind, keeping promises, or being honest, or voting for the social democrats—they care for these things only indirectly. What they really care about is that they do the right thing, or that they are moral. Smith labels the difference between direct moral concern (and motivation) and indirect moral concern (and motivation) as *de re* and *de dicto,* respectively.

Externalism, according to Smith, implies that moral concern and moral motivation is *de dicto*, not *de re*.

There are two problems with this view of motivation according to Smith. First, he argues that it is evidence of a kind of *moral fetishism*, where people have only a single and overriding desire to do the right thing. These people are obsessive about being moral, but they are not at all concerned with the actual content of morality (e.g., be kind, vote for the social democrats). This, he maintains, is a supreme vice, not a virtue, because, at bottom, moral concern and moral motivation is wholly self-regarding; people care that they themselves do the right thing without being directly concerned for others. Secondly, on the externalist's conception of motivation, people would be *alienated* from their moral judgments. There is necessarily one further thought to get from the judgment to the motivation, but in ordinary moral thought, this is not the case. Ordinarily, a person is moved directly by his moral judgments, without any intervening steps. Therefore, the externalist explanation of moral motivation entails a vicious and alienated view of moral motivation that is quite inconsistent with the character and dispositions of the good and strong-willed person. This, according to Smith, is a straightforward *reductio* of externalism, and therefore internalism is the correct account of the reliable connection between a moral judgment and the motivation to act consistently with it.

However, there is a serious problem with Smith's argument in that it assumes that because internalism provides a better account of moral motivation in some cases than externalism, that it therefore provides the right account of moral motivation in all cases; that is, it assumes that there is one single true spring of moral motivation. But

this is a mistake, because there is no reason to think, at least *a priori*, that all moral judgments motivate in the same way. That is an empirical question (in the sense of being *a posteriori*), and there are good reasons to think that *some* moral judgments motivate only *de dicto*, not *de re*—that people are motivated to act consistently with them only because they want to do the right thing, not because they are motivated directly by the content of their moral judgment.

In response Smith might say that in such cases a person is guilty of a vicious sort of moral fetishism. But this is a really odd sort of criticism. It is generally a good thing to be concerned with morality, as such, and to care about doing the right thing and about being a good person. Imagine being in a position where one has just had to muster the motivation to act consistently with a moral judgment because one wants to do the right thing, struggling against many competing desires and motives, and having done it just because it was the right thing to do, and then being met by Smith and told that one's seemingly courageous and strong-willed act was evidence of an unhealthy and undue obsession with morality. Such a criticism would be downright bizarre, and utterly misplaced. There is something odd about accusing people who demonstrate a remarkable strength of will in this way of being viciously morally deformed. Sometimes *de dicto* motivation is precisely what constitutes the motivational structure of a morally serious and well-formed agent.

And Smith's case of coming to a change of mind about the social democrats itself provides just the right sort of case where *de dicto* motivation is likely, and perhaps even morally preferable. If a person really does come to change his most fundamental values—values that underlie all other sorts of evaluative judgments and

that are deeply connected to that person's will, dispositions and character—is it plausible to think that such a person can turn his back on such values automatically and effortlessly and whole-heartedly embrace his new ones? Is this really how a good and strong-willed person would and *should* respond, as Smith argues? It is extremely doubtful. In fact, such a person would be quite puzzling, because ordinary people cannot so easily adopt a whole new set of values that contradict ones they previously held. Changing one's fundamental moral orientation and motivational dispositions is not as simple as flipping a switch. Motivational dispositions take time to change, and doing so is far from easy. And before such changes take place sometimes the only way to muster the motivation to act consistently with a moral judgment is to do because it is the right thing to do.

By attempting to fit all moral motivation in the same explanatory box, Smith provides an implausible picture of the actual complexities of moral motivation. This is a serious explanatory worry for internalism, and it gives us reason to ask whether externalism might fare any better. It does not.

Externalists claim that *all* moral motivation is realized by some desire or other conative attitude under a moral mode of presentation, such as the desire to be moral or the desire to do the right thing (Svavarsdottir, 1999). But the externalist's view of moral motivation raises a serious worry in that it gets at least one feature of moral motivation quite wrong, namely, that there are certainly some moral judgments that motivate directly. For example, if Jones judges that he should keep his promise to meet a friend, he is straightaway motivated to meet his friend. The motivation is not mediated by some intervening desire to be moral or do to the right thing—the

judgment follows immediately from the judgment that he should keep his promise. Jones is not concerned in this case with doing the right thing, but in keeping his promise. That is, he has a direct motive (*de re*) to keep his promise to meet his friend, not an indirect motive to do the right thing (*de dicto*). Of course, it is possible to claim that there really is an intervening desire in this direct motivation, and that this intervening desire gives rise to a direct concern for Jones to keep his promise, but in this case, Smith is right—it is one step too many. There is no reason to maintain that such a desire is implicated in every instance of moral motivation except the desire to save a simplistic picture of moral motivation. But motivational dispositions are complex, even in the moral case, and it is false to the facts to maintain that every instance of moral motivation is facilitated by a single desire.[3]

What is needed to fully capture the facts of moral motivation is a psychological picture that can explain how some moral judgments motivate directly while other moral judgments motivate only indirectly—how some moral judgments are closely connected with the will, while others are at a further remove (Railton, 2006). Sadler (2003) makes such a proposal when she writes:

> Logically, if externalism is simply the denial of the internalist thesis, and if internalism is not true (or is at least unpersuasive), that would seem to make my position externalist. However, something more substantive needs to be said on this point. After all, there seems to be room for something like a middle-ground: the internalist thesis is not

[3] Svavarsdottir (1999) argues that one way around this problem is to posit a set of moral desires, as opposed to a single desire to be moral; such as a desire to be kind, a desire to keep one's promises, etc, which can account for direct moral concern and motivation. This is a nice solution, and consistent with the argument I develop.

> always true, or true insofar as it is regarded as a universal generalization that all rational agents are always motivated by their moral judgments, though it may sometimes be true of some agents. Although anything shy of endorsing the universal truth of the internalist thesis will count as externalism, such a 'middle ground' position seems to make the internalism/externalism debate appear less concerned with purely conceptual questions and more amenable to empirical observations (pg. 73).

One very straightforward way to make good on Sadler's proposal is to reject the Unity of Kind Thesis. Giving an account of moral motivation while operating under the constraints of the Unity of Kind Thesis leads quite naturally to the view that if moral judgments necessarily motivate, then both intuitive and deliberative moral judgments necessarily motivate in the same way; and if moral judgments do not necessarily motivate, then both intuitive and deliberative moral judgments depend on the presence of a suitable desire to motivate. The problem with the Unity of Kind Thesis with respect to moral motivation, though, is that it makes it quite difficult to explain how some moral judgments motivate directly, while others motivate only indirectly. No universal account of moral motivation can fit both of these cases equally well. Rejecting the Unity of Kind Thesis provides a very straightforward way of resolving this impasse.

According to the Two Kinds Hypothesis, intuitive and deliberative moral judgments can motivate differently because they are different psychological kinds

that play different roles with respect to moral thinking and acting. Intuitive moral judgments are, in some way, emotional judgments, and thus they motivate directly in the absence of any generalized motivational problems, such as tiredness, depression or apathy (see, Stocker, 1979 for a discussion). Deliberative moral judgments, on the other hand, are more like ordinary beliefs, and they motivate indirectly via a desire, such as, perhaps, the desire to do the right thing, or the desire to be moral. To put the motivational difference between intuitive and deliberative moral judgments in terms of Smith's distinction between *de re* and *de dicto* motivation, intuitive moral judgments motivate *de re* in the ordinary person, while deliberative moral judgments motivate *de dicto*. This is why people often *fail* to act consistently with their deliberative moral judgments. It is not because they are not genuine moral judgments; it is because they do not motivate in the same direct way as intuitive moral judgments, and acting on them sometimes takes tremendous effort, especially if it conflicts with one's settled motivational dispositions and habits of character. This is why some people try to suppress their deliberative moral judgments when acting, or ignore them, or otherwise put them out of mind, because this can be an easier strategy than acting consistently with them.

This explanation of moral motivation provided by the Two Kinds Hypothesis provides a much better account of Smith's example of the person changing his fundamental values, and can show what is so bizarre in Smith's example of the person who comes to change his fundamental values and is immediately motivated to act consistently with them. Here we are to picture someone whose deliberation does not produce a new intuitive moral judgment: he does not reframe the issues or objects

of evaluation in such a way to see that they actually fall under some concept that the person already judges to be wrong or something like that. That sort of case does not count as a change in *fundamental* values at all. To count as a change in fundamental values we have to imagine someone who reasons himself to a whole new set of values.

So, imagine someone who previously considered justice to be a fundamental moral value, and has now become convinced of the truth of utilitarianism, specifically hedonic rule-utilitarianism, and now has to decide who to vote for. His whole life he has identified with the libertarians, and has viewed libertarian politicians as heroes to the cause of justice. But now, having changed his fundamental values, he sees that voting for the social democrats is morally required, because their policies will provide the greatest pleasure to the greatest number. He judges that voting for the social democrats is the right thing to do. And yet, he is not straightaway motivated to vote for the social democrats—and when he gets inside the voting booth he is deeply conflicted about whether he should actually do it. He feels that doing so would, in some ways, be an act of betrayal to the cause of justice. In the end the only way he is able to vote for the social democrats is that he thinks it important to live by his moral convictions, regardless of his personal feelings regret.

Filling out Smith's case in this way, the person typifies moral courage. The conflict he experiences is understandable, but in the end, this person is good and strong-willed because he is able to act on his carefully considered deliberative moral judgment, even though doing so is difficult, and it very much conflicts with his settled dispositions and habits of character. This is what moral courage really looks like, and

it is this deep sort of internal moral conflict and competing motivations that the Two Kinds Hypothesis explains. This person is finally and sufficiently motivated to do the right thing just because it is the right thing to do. By contrast, the character Smith asks us to imagine is wooden, lacking in depth, and lacking in just this sort of moral courage that typifies the good and strong-willed person.

The Two Kinds Hypothesis provides a straightforward way of explaining how it is some moral judgments motivate differently than others, and in a way that captures the actual complexities of moral motivation. Moreover, the Two Kinds Hypothesis can thread the "middle way" called for by Sadler, capturing what is true of internalism, and what is true of externalism, and providing a rapprochement between the two positions. Moreover, it shows that internalists and externalists, by implicitly accepting the Unity of Kind Thesis, are often talking at cross-purposes, because they each focus on different kinds of moral judgments. The best way to proceed is by indexing universal motivational claims with respect to moral judgments to either intuitive or deliberative moral judgments.

2. Generalism and Particularism

A second metaethical debate that the Two Kinds Hypothesis can help reorganize and reconceptualize is the debate between moral generalists and moral particularists. In some ways, however, it is quite difficult to characterize the positions of generalists and particularists, except to say that generalists are the opponents of particularists, and *vice versa*. The central questions at issue in this debate, however, are simple enough: what role *do* moral principles have in moral thinking; and what role *should* moral principles have in moral thinking? It is tempting to say that

particularists answer both of these questions by saying that moral principles have no role, and should have no role, in moral thinking, however, it turns out that answering these questions is not a simple matter. McKeever and Ridge, for example, distinguish between six possible conceptions of a moral principle, five different ways a moral principle can be rejected, and end up listing 120 different combinatorial possible particularists positions (McKeever & Ridge, 2006). This is not a promising start, because it is not possible to claim that particularism is *a* view that can be given a single characterization. But, even if there are 120 different possible particularists positions, only few particularist positions have been staked out in the literature. Among the most influential are Little and Lance's view that particularism rejects "unhedged" moral principles (Lance & Little, 2004, 2005; Little, 2000),[4] Hotlon's view that particularism is the view that there are moral principles, but that no suitable provision of them suffices to cover the entire moral terrain (Holton, 2002), and Dancy's view, which is that "the possibility of moral thought and judgment does not depend on the provision of a suitable supply of moral principles" (Dancy, 2004, p. 7).

Even within these three particularist views, it is not clear whether particularism should first be understood as an epistemological claim, or a metaphysical one. Indeed, even individual particularists are not clear always clear about which kind of claim they intend to be making. Dancy, for example, writes that he "used to think that particularism was a position in moral epistemology," but now he thinks it is a position in moral metaphysics (Dancy, 2004, p. 140). This sort of

[4] "Unhedged" moral principles are moral principles that do not admit of an exception.

confusion is partly what makes the debate between particularists and generalists notoriously obscure.

In this discussion, however, I shall principally focus on particularists' epistemological claim, which can be characterized generally as the claim that moral principles play little or no role in moral thinking, and that they play little or no role in the production of moral judgments. Generalism can then be characterized as the rejection of this epistemological claim.[5] Generalists, allow that moral judgments can be arrived at without consciously considering moral principles, but they argue that moral principles are necessary either to explain the reliability of an agent's moral judgments from context to context (Jackson, Pettit, & Smith, 2000), moral justification, or our practices with respect to moral disagreements (Sinnott-Armstrong, 1999).

Again, the Two Kinds Hypothesis can do some work here in helping to reorganize and reconceptualize the debate between moral particularists and generalists, because the epistemological debate between generalists and particularists involves a psychological claim with respect to role of moral principles in moral reasoning. It is with respect to this claim that the Two Kinds Hypothesis can do some work. Moreover, by helping reorganize and reconceptualize this psychological dispute, the Two Kinds Hypothesis can also bear indirectly on the metaphysical dispute between particularists and generalists, because a principal support for the

[5] Of course, there are many generalists whose positions are principally characterized as the rejection of the particularists' metaphysical claim that the moral domain is not exhausted by a suitable provision of moral principles. With respect to the metaphysical claim, Väyrynen (2006), for example, argues that the structure of moral domain is fully exhausted by a suitable provision of "hedged moral principles", while others argue that Ross's theory of *prima facie* duties (Ross, 1930) provides a suitable list of moral principles that exhaust moral domain (Crisp, 2000; Hooker, 2000).

metaphysical claims of particularists is their understanding of the psychology of moral judging, namely, that mature moral agents can arrive at moral judgments without reference to consciously accessible general moral principles. If it can be shown that this observation is entirely consistent with the possibility that mature moral agents employ moral principles of some kind in their moral thinking, then at the very least it will show that the metaphysical claim of particularism is under-motivated.

Here is a suggestion that can be worked into clearer focus: the psychological observations with respect to moral judging that inspire particularists' arguments are focused almost exclusively on intuitive moral judging, and the psychological observations with respect to moral judging that motivate generalist arguments are focused almost exclusively on deliberative moral judging. I say almost exclusively here, because particularists attempt to accommodate *some* features of deliberative moral judging (that it can be used in education), and generalists try to accommodate *some* features of intuitive moral judging (that moral judgments can be context sensitive), but they both run into problems explaining *all* features of moral judging because neither pays attention to the distinction between intuitive and deliberative moral judging, which leads them to treat all moral judgments as a single psychological kind. However, since both particularists and generalists attempt to accommodate features of both intuitive and deliberative moral judging, my initial suggestion should be refined to the claim that the picture of moral judging of particularism is primarily motivated by considerations with respect to intuitive moral

judging, and the picture of moral judging of generalism is primarily motivated by considerations with respect to deliberative moral judging.

This suggestion can be fleshed out by looking at the explananda particularists and generalists take it their pictures of moral judging are meant to explain. Particularists seek to explain how it is agents are able to make quick and automatic moral judgments that strike them almost as perceptions without any conscious reasoning, and sometimes without any reasons that could be cited for the judgment. For example, in a typical particularist passage Little writes:

> According to particularists, we can come to discern or interpret the moral nature of specific actions or individuals by exercising a sensitivity—a sensitivity that is perhaps analogous to a perceptual capacity, but is perhaps just a species of the more familiar 'faculty' we use to apprehend that something is a table, namely, the capacity or skill to apply concepts correctly (2000, pg. 292).

In some cases, of course, moral judgments do have a very similar phenomenology to perceptual judgments, but this phenomenology only characterizes intuitive moral judgments. Thus, particularists' epistemological claims are primarily motivated by observations with respect to intuitive moral judging. Generalists, on the other hand, do not deny these explananda, but seek to explain them within a picture of moral judging that can explain the fact that people sometimes do reason themselves to a moral judgment using moral principles, and that the use of moral principles are

pervasive in our ordinary practices of justification, especially in the context of moral disagreement.

Sinnott-Armstrong (1999), for example, objects to particularism because it gets the facts of moral disagreement wrong. When two people disagree in their moral judgments with respect to some action or person, they do not simply try to get the other person to be more sensitive or to look harder; rather, they appeal to moral principles to help settle the matter between them. Similarly, if a person experiences a moral conflict, she appeals to moral principles to help her resolve it. Research with respect to moral dumbfounding reveals that sometimes such principles may simply be post hoc confabulations, but it is also possible that such principles can be the basis of a person's moral judgment, such as one a person might appeal to in settling in internal moral conflict.

Even some particularists complain that, with respect to our practices of moral justification, particularists have overstated their case (Little, 2000). When it comes to justifying our moral judgments people often make appeals to general moral presumptions, such as "stabbing is presumptive of cruelty," and the usefulness of such general moral presumptions implies the possibility that some moral judgments are arrived at by employing them in moral deliberation. Little is a particularist, but she is concerned that particularists undercut the plausibility of their epistemological claims by rejecting any useful role for moral principles in the moral thinking of mature moral judgers. Indeed, she argues, that inasmuch as particularists see the possibility that general moral principles can be pedagogically useful, they implicitly endorse the usefulness of employing them in moral deliberation; principles are useful

precisely because, *ceteris paribus*, such presumptions lead to appropriate deliberative moral judgments. If our practices with respect to moral justification rely on the possibility of arriving at a moral judgment by inference from a general moral presumption, then, argues Little, moral justification requires deliberative moral judging. In deliberative moral judging a person can consider moral presumptions, moral principles, and explicit beliefs to produce a deliberative moral judgment. More importantly, in deliberative moral judging a person can determine whether his or her moral judgments are actually appropriate, not just that they appear appropriate.

What these arguments reveal is that the primary motivation for generalism, or at least the primary problem for particularism, are certain facts with respect to deliberative moral judging. Generalists are at least usually aware of the insights of particularists with respect to psychology of moral judging, but the only way generalists can explain them is by relying on some form of moral expertise. Hare (Hare, 1981), for example, argues that the kind of facts with respect to moral judging that particularists point to can be explained by the internalization of general principles.[6] However, moral expertise views cannot fully account for the capacity of intuitive moral judging, so generalist pictures of moral judging cannot explain all features of moral judging. The important point here, though, is that, the disparate pictures of moral judging of particularists and generalists are the result of particularists focusing on intuitive moral judgments, and generalists focusing on deliberative moral judgments.

This is where the Two Kinds Hypothesis can help reorganize and reconceptualize the dispute between particularists and generalists with respect to the

[6] See Blum (2000) for a particularist critique of Hare's analysis.

particularist's epistemological claim, because it offers a framework of the distinction between intuitive and deliberative moral judging and judgment that fully captures the insights and concerns particularists and generalists, respectively, but it does so by rejecting the assumption that it is possible to explain all features of moral judging by a single psychological process, and by rejecting the assumption that all moral judgments are a single psychological kind. Both particularists and generalists implicitly assume the truth of the Unity of Process Thesis and the Unity of Kind Thesis, and these assumptions are what really underlie their dispute. However, I have already shown that there are very good reasons for thinking that intuitive and deliberative moral judging are different, functionally distinct and causally independent psychological processes, and thus that Unity of Process Thesis is false. One consequence of rejecting the Unity of Process Thesis is that universal claims with respect to moral judging need to be indexed to either intuitive or deliberative moral judging.

Indexing the universal claims of particularists and generalists is fairly straightforward: particularists' epistemic claims with respect to moral judging and judgment should be indexed to intuitive moral judging, and generalists' epistemic claims should be indexed to deliberative moral judging and judgment. If this is right, then the epistemic dispute between particularists and generalists simply dissolves, because they are each referring to a different kind of psychological process for moral judging that produce different kinds of moral judgments. Once their claims are indexed, particularists and generalists are simply talking past each other, at least with respect their epistemic claims.

When it comes to the metaphysical claim of particularists, however, the Two Kinds Hypothesis cannot directly help. But, what the Two Kinds Hypothesis can do is at least provide some reasons for resisting the move from moral psychology to moral metaphysics. The particularist's claim that the structure of moral domain cannot be exhausted by a suitable provision of moral principle is primarily motivated by the observation that the moral judgments of mature moral agents cannot be codified by any possible set of moral principles. It is always possible to imagine cases where any proposed moral principle might admit of exceptions, reversals, or the like in agent's actual moral judgments. Dancy, in particular, focuses on examples where highly intuitive moral principles such as "do not lie" are prone to reversals—that is, in some cases, the fact that something is a lie is a reason *to* do it. For example, if telling a trivial lie is the only way to save someone's life, then the fact that some statement is a lie is a reason to say it. According to Dancy and other particularists, reversals like this show that agents' moral judgments are sensitive to contextual features in a way that moral principles are not,[7] and this is taken as evidence that the moral domain cannot be exhausted by a suitable provision of moral principles because there is no clear supervenience relation between the non-moral facts that moral principles pick out and the moral judgments of mature moral agents.

One thing to notice about this argument is that the pressure to accept the conclusion that the moral domain cannot be fully exhausted by a suitable provision of moral principles is provided by the observation that the psychological processes of

[7] One might think that the reversal we see in the case of telling a trivial lie in order to save a life indicates a lexical ordering of moral principles, such that telling a lie is prima facie wrong, but when it conflicts with a lexically higher moral principle, such as saving a life, then it is not wrong all things considered. Dancy argues that this is the wrong analysis, because it assumes that "complete reasons" include contributory facts.

intuitive moral judging that mediate the movement in a person's mind from the apprehension of some non-moral facts to a moral judgment is not fully specifiable in terms of moral principles. This is a psychological claim, and one that is readily accounted for by the Two Kinds Hypothesis. Regardless of the best way to understand the System 1-type processes of intuitive moral judging, nearly all extant theories of intuitive moral judging reject the idea that specifiable moral principles play any kind of causal role in the production of intuitive moral judgments. Rather, many argue that System 1-type intuitive moral judging relies on prototypes, exemplars, or narratives (Sripada & Stich, 2006), connectionist networks (Churchland, 1996), pattern recognition (Sterelny, 2008), heuristics (Gigerenzer, 2008), or a function in intension (Dwyer, 2009). None of these could be fully specified by a set of moral principles.

Thus, there is a perfectly good psychological explanation for why the intuitive moral judgments of mature agents cannot be fully specified by a suitable provision of moral principles. This psychological observation, however, does nothing to support the metaphysical claim that the moral domain therefore cannot be fully exhausted by a suitable provision of moral principles. It is at least possible that someone could discover through deliberation that any number of possible moral principles really do comprehensively describe the moral domain, and moreover, that at least some of the intuitive moral judgments of mature moral agents are inappropriate. Perhaps utilitarianism is right, and any action that maximizes happiness (whatever we take that to mean) really is always the right action. Or, perhaps Kant's Formula of Humanity is correct, and any action that violates a person's humanity is always

wrong. Or, perhaps a pluralistic account such as Ross's comprehensively describes the moral domain. The point here is not to defend any of these moral theories, but rather to show that once the claims of particularists are properly indexed to intuitive moral judging, there is no reason to think that their psychological observations have any metaphysical consequences, because it could well be that the actual boundaries of the moral domain is, in principle, specifiable after sufficient deliberation.

3. Conclusion

One significant means of evaluating any hypothesis is to assess whether it provides the best explanation of all the available data of the target phenomena it is meant to explain. Over the past five chapters I have argued that the Two Kinds Hypothesis provides the best explanation of recent empirical research with respect to moral judging and judgments, better captures a range of our complex moral practices, and better captures central aspects of our lived moral experiences than other contemporary models of our moral judging. This provides very strong support for the Two Kinds Hypothesis, and gives very good reasons for thinking that it provides the right framework for moral judging.

Another significant means of evaluating a hypothesis is whether it illuminates issues that are not themselves part of the target phenomena it is meant to explain. If it does, then the hypothesis is *fruitful*. I have argued in this chapter that the Two Kinds Hypothesis can be usefully applied to at least two debates in metaethics to help reorganize and reconceptualize them to provide a way forward in some seemingly intractable disputes. It does not decide any of these issues, nor could it, because metaethical debates involve metaphysical, conceptual, empirical, and psychological

claims, but in providing a way forward, the Two Kinds Hypothesis is a fruitful hypothesis as well as the best explanatory framework for our ordinary capacities for moral judging. Again, this is less than conclusive proof for the truth of the Two Kinds Hypothesis, but it does show that the Two Kinds Hypothesis is a powerful framework for understanding our capacities for moral judging, and should serve as a basis for future empirical and philosophical research.

Bibliography

Aristotle. (1999). *Nicomachean Ethics* (Terence Irwin, Trans. 2nd ed.). Indianapolis: Hackett Publishing Company.

Ayer, A. J. (1952). *Language, Truth and Logic*. New York: Dover Publications.

Bargh, John A., & Chartrand, Tanya L. (1999). The Unberable Automaticity of Being. *American Psychologist, 54*, 462-479.

Bargh, John A., Chen, M., & Burrows, L. (1996). The Automaticity of Social Behavior: Direct Effects of Construct and Stereotype Activation on Action. *Journal of Personality and Social Psychology, 71*, 230-244.

Baron, Jonathan. (1985). *Rationality and Intelligence*. New York: Cambridge University Press.

Baron, Jonathan. (2003). Myside Bias in Thinking About Abortion. *Unpublished manuscript*.

Berker, Selim. (2009). The Normative Insignificance of Neuroscience. *Philosophy and Public Affairs, 37*(4), 293-329.

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., Grodd, W., & Flor, H. (2005). Deficient Fear Conditioning in Psychopathy A Functional Magnetic Resonance Imaging Study. *Archives of General Psychiatry, 62*(7), 799-805.

Blair, James. (2009). How Do The Moralities Develop? *Center for Children, Relationships, and Culture & Developmental Science Field Committee Co-Sponsored Colloquium Series*.

Blair, R. J. R. (1995). A Cognitive Developmental Approach to Morality: Investigating the Psychopath. *Cognition, 57*, 1-29.

Blair, R. J. R. (1999). Responsiveness to Distress Cues in the Child with Psychopathic Tendencies. *Personality and Individual Differences, 27*(1), 135-145.

Blair, R. J. R., Jones, L., Clark, F., & Smith, M. (1995). Is the Psychopath "Morally Insane"? *Personality and Individual Differences, 19*, 741-752.

Blair, R.J.R. (1997). Affect and the Moral-Conventional Distinction. *Journal of Moral Education, 26*(2), 187-196.

Blair, R.J.R., Monson, Jey, & Frederickson, Norah. (2001). Moral Reasoning and Conduct Problems in Children with Emotional and Behavioural Difficulties. *Personality and Individual Differences, 31*, 799-811.

Bloom, Paul. (2007). *The Moral Circle*. Paper presented at the Johns Hopkins Evolution, Cognition, and Culture Project, Johns Hopkins University.

Bloom, Paul. (2010). How Do Morals Change? *Nature, 464*, 490.

Blum, Lawrence. (1980). *Friendship, Altruism, and Morality*. New York: Routledge & Kegan Paul Books.

Blum, Lawrence. (1991). Moral Perception and Particularity. *Ethics, 101*, 701-725.

Blum, Lawrence. (2000). Against Deriving Particularity. In Brad Hooker and Margaret Little (Eds.), *Moral Particularism*. New York: Oxford University Press.

Bodenhausen, Galen V. (1990). Stereotypes as Judgmental Heuristics: Evidence of Circadian Variations in Discrimination. *Psychological Science, 1* (5), 319-322.

Brenner, Lyle, Koehler, Derek, & Rottenstreich, Yuval. (2002). Remarks on Support Theory: Recent Advances and Future Directions. In Thomas Gilovich, Dale Griffin & Daniel Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 489-509): Cambridge University Press.

Brink, David O. (1989). *Moral Realism and the Foundations of Ethics*: Cambridge University Press.

Carruthers, Peter. (2009). An Architecture for Dual Reasoning. In Jonathan Evans & Keith Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 109-128). New York: Oxford University Press.

Cholbi, Michael. (2006a). Belief Attribution and the Falsification of Motive Internalism. *Philosophical Psychology, 19*(5), 607-616.

Cholbi, Michael. (2006b). Moral Belief Attribution: a Reply to Roskies. *Philosophical Psychology, 19*(5), 629-638.

Chomsky, Noam. (1980). *Rules and Representations*. New York: Columbia University Press.

Churchland, Paul M. (1996). The Neural Representation of the Social World. In Larry May, Marilyn Friedman & Andy Clark (Eds.), *Mind and Morals: Essays on Cognitive Science and Ethics* (pp. 91-108). Cambridge, MA: MIT Press.

Cleckley, Hervey M. (1964). *The Mask of Sanity: An Attempt to Clarify Some Issues About the So-Called Psychopathic Personality* (4th ed.). St. Louis: C.V. Mosby Co.

Colby, A., Kohlberg, L., Gibbs, J., & Lieberman, M. (1983). A Longitudinal Study of Moral Judgment. *Monographs of the Society for Research in Child Development, 48*(1-2), 1-96.

Cosmides, Leda. (1989). The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task. *Cognition, 31*(3), 187-276.

Craigie, Jillian. (2011). Thinking and Feeling: Moral Deliberation in a Dual-Process Framework. *Philosophical Psychology, 24*(1), 53-71.

Crisp, Roger. (2000). Particularizing Particularism. In Brad Hooker & Margaret Little (Eds.), *Moral Particularism* (pp. 23-47). New York: Oxford University Press.

Cummins, Robert. (1975). Functional Analysis. *The Journal of Philosophy, 72*(20), 741-765.

Cummins, Robert. (2000). "How Does it Work?" versus, "What Are the Laws?": Two Conceptions of Psychological Explanation. In Frank C. Keil & Robert A. Wilson (Eds.), *Explanation and Cognition* (pp. 117-144). Cambridge, MA: MIT Press.

Cushman, Fiery, Young, Liane, & Hauser, Marc D. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgments: Testing Three Principles of Harm. *Psychological Science, 17*(12), 1082-1089.

D'Arms, Justin, & Jacobson, Daniel. (2000a). The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophy and Phenomenological Research, 61*(1), 65-90.

D'Arms, Justin, & Jacobson, Daniel. (2000b). Sentiment and Value. *Ethics, 110*(4), 722.
Damasio, Antonio. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. New York: G.P. Putnam & Sons.
Dancy, Jonathan. (2000). *Practical Reality*. New York: Oxford University Press.
Dancy, Jonathan. (2004). *Ethics Without Principles*. New York: Oxford University Press.
Daniels, Norman. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy, 76*(5), 256-282.
Darwall, Stephen, Gibbard, Allan, & Railton, Peter (Eds.). (1997). *Moral Discourse & Ethics*. New York: Oxford University Press.
de Sousa, Ronald. (2003). Emotion. In Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Vol. Spring 2003). Stanford, CA: The Metaphysics Research Lab.
Deigh, John. (1995). Empathy and Universalizability. *Ethics, 105* (4), 743-763.
Dershowitz, Alan M. (2002). *Why Terrorism Works: Understanding the Threat, Responding to the Challenge*. New Haven, CT: Yale University Press.
Doris, John. (2005). *Lack of Character*. New York: Cambridge University Press.
Dreier, James. (1990). Internalism and Speaker Relativism. *Ethics, 101*(1), 6-26.
Dwyer, Susan. (1999). Moral Competence. In Kumiko Murasugi & Robert Stainton (Eds.), *Philosophy and Linguistics*. Boulder, CO: Westview Press.
Dwyer, Susan. (2006). How Good is the Linguistic Analogy? In Peter Carruthers, Stephen Laurence & Stephen Stich (Eds.), *The Innate Mind Volume 2: Culture and Cognition* (pp. 237-256). New York: Oxford University Press.
Dwyer, Susan. (2009). Moral Dumbfounding and the Linguistic Analogy: Methodological Implications for the Study of Moral Judgment. *Mind & Language, 24*(3), 274-296.
Evans, Jonathan. (2008a). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology, 59*(6), 1-24.
Evans, Jonathan. (2008b). How Many Dual-Process Theories Do We Need? One, Two or Many? In Jonathan Evans & Keith Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 33-54). New York: Oxford University Press.
Evans, Jonathan, & Over, David. (1996). *Rationality and Reasoning*. New York: Psychology Press.
Evans, Jonathan, & Over, David. (1999). *Rationality and Reasoning*. New York: Psychology Press.
Fessler, Daniel M.T., Arguello, Alexander P., Mekdara, Jeannette M., & Macias, Ramon. (2003). Disgust Sensitivity and Meat Consumption: A Test of an Emotivist Account of Moral Vegetarianism. *Appetite, 41*(1), 31-41.
Fine, Cordelia. (2006). Is the Emotional Dog Wagging Its Tail, or Chasing It? *Philosophical Explorations, 9*(1), 83-98.
Finlay, Stephen, & Schroeder, Mark. (2008). Reasons for Action: Internal vs. External. In Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Vol. Spring 2009). Stanford, CA: The Metaphysics Research Lab.

Finucane, Melissa, Alhakami, Ali, Slovic, Paul, & Johnson, Stephen. (2000). The Affect Heuristic in Judgments of Risks and Benefits. *Journal of Behavioral Decision Making, 13*, 1-17.
Flanagan, Owen J. (1991). *Varieties of Moral Personality*. Cambridge, MA: Harvard University Press.
Foot, Philippa. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*(5).
Frankfurt, Harry. (1988). Rationality and the Unthinkable *The Importance of What We Care About*. New York: Cambridge University Press.
Gazzaniga, Michael. (1995). Consciousness and the Cerebral Hemispheres. In Michael Gazzaniga (Ed.), *The Cognitive Neurosciences*: MIT Press.
Gewirth, Alan. (1988). Ethical Universalism and Particularity. *Journal of Philosophy, 85*(6), 283-302.
Gibbard, Allan. (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.
Gigerenzer, Gert. (2008). Moral Intuition = Fast and Frugal Heurstics? In Walter Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cogntive Science of Morality: Intuition and Diversity*. Cambridge, MA: MIT Press.
Gilbert, Daniel T. (1999). What the Mind's Not. In Shelly Chaiken & Yaacov Trope (Eds.), *Dual Process Theories in Social Psychology* (pp. 3-11). New York: Guilford.
Gilbert, Daniel T. (2002). Inferential Correction. In Thomas Gilovich, Dale Griffin & Daniel Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 167-184). New York: Cambridge University Press.
Gilligan, Carol. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.
Greene, Joshua. (2005). Cognitive Neuroscience and the Structure of the Moral Mind. In Peter Carruthers, Stephen Laurence & Stephen Stich (Eds.), *The Innate Mind: Structure and Contents* (pp. 338-352). New York: Oxford University Press.
Greene, Joshua. (2007). Why are VMPFC Patients More Utilitarian? A Dual-Process Theory of Moral Judgment Explains. *Trends in Cognitive Sciences, 11*(8), 322-323.
Greene, Joshua. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*: MIT Press.
Greene, Joshua, & Haidt, Jonathan. (2002). How (and Where) Does Moral Judgment Work? *Trends in Cognitive Sciences, 6*(12), 517-523.
Greene, Joshua, Nystrom, Leigh E., Engell, Andrew D., Darley, John M., & Cohen, Jonathan D. (2004). The Neural Basis of Cognitive Conflict and Control in Moral Judgment. *Neuron, 44*, 389-400.
Greene, Joshua, Sommerville, R. Brian, Nystrom, Leigh E., Darley, John M., & Cohen, Jonathan D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science, 293*, 2105-2108.
Greenspan, Patricia. (1988). *Emotions and Reasons*. New York: Routledge.

Haidt, J., & Joseph, Craig. (2007). The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Cultural-Specific Virtues, and Perhaps Even Modules In Peter Carruthers, Stephen Laurence & Stephen Stich (Eds.), *The Innate Mind Volume 3: Foundations and the Future* (pp. 367-392). New York: Oxford University Press.

Haidt, Jonathan. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review, 108*(4), 814-834.

Haidt, Jonathan, & Bjorklund, Frederik. (2008a). Social Intuitionist Answer Six Questions. In Walter Sinnott-Armstrong (Ed.), *Moral Psychology Vol. 2: The Cognitive Science of Morality: Intuition and Diversity* (pp. 181-218). Cambridge, MA: MIT Press.

Haidt, Jonathan, & Bjorklund, Frederik. (2008b). Social Intuitionists Reason, In Conversation. In Walter Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity* (pp. 241-254). Cambridge, MA: MIT Press.

Haidt, Jonathan, Bjorklund, Frederik, & Murphy, Scott. (2000). Moral Dumbfounding: When Intuition Finds No Reason. *Unpublished manuscript*.

Haidt, Jonathan, & Hersh, M.A. (2001). Sexual Morality: The Cultures and Reasons of Liberals and Conservatives. *Journal of Applied Social Psychology, 31*, 191-221.

Haidt, Jonathan, & Joseph, Craig. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus, 133*.

Haidt, Jonathan, Koller, Silvia H., & Dias, Maria G. (1993). Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog? *Journal of Personality and Social Psychology, 65*(4), 613-628.

Hare, R.M. (1952). *The Language of Morals*. New York: Oxford University Press.

Hare, R.M. (1981). *Moral Thinking: Its Levels, Method, and Point*. New York: Oxford University Press.

Harman, Gilbert. (1975). Moral Relativism Defended. *Philosophical Review, 84*(3), 3-22.

Harman, Gilbert. (1997). Ethics and Observation. In Stephen Darwall, Allan Gibbard & Peter Railton (Eds.), *Moral Discourse & Practice* (pp. 83-88). New York: Oxford University Press.

Harman, Gilbert. (1999). Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society, 99*, 315-331.

Hauser, Marc D. (2006). *Moral Minds*. New York: Ecco.

Hauser, Marc D., Cushman, Fiery, Young, Liane, Jin, R. Kang-Xing, & Mikhail, John. (2007). A Dissocation Between Moral Judgments and Justification. *Mind & Language, 22*(1), 1-21.

Hauser, Marc D., Young, Liane, & Cushman, Fiery. (2008). Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In Walter Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2: The Cognitive Science of Morality: Intuition and Diversity* (pp. 107-144). Cambridge, MA: MIT Press.

Held, Virginia. (1996). Whose Agenda? Ethics versus Cognitive Science. In Larry May, Marilyn Friedman & Andy Clark (Eds.), *Mind and Morals: Essays on Ethics and Cognitive Science* (pp. 69-88). Cambridge, MA: MIT Press.
Herman, Barbara. (1993). *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.
Holton, Richard. (2002). Principles and Particularism. *Proceedings of the Aristotelian Society, Supplemental*, 169-210.
Hooker, Brad. (2000). Moral Particularism: Wrong and Bad. In Brad Hooker & Margaret Little (Eds.), *Moral Particularism* (pp. 1-22). New York: Oxford University Press.
Horgan, Terry, & Timmons, Mark. (2007). Morphological Rationalism and the Psychology of Moral Judgment. *Ethical Theory and Moral Practice, 10*, 279-295.
Huebner, Bryce, Dwyer, Susan, & Hauser, Marc D. (2009). The Role of Emotion in Moral Psychology. *Trends in Cognitive Sciences, 13*(1), 1-6.
Hughes, Gerard. (2003). *Routledge Philosophy Guidebook to Aristotle: On Ethics*. New York: Routledge Press.
Hume, David. (1739/1978). *A Treatise of Human Nature*. New York: Oxford University Press.
Hussar, Karen M., & Harris, Paul L. (2010). Children Who Choose Not to Eat Meat: A Study in Early Moral Decision-Making. *Social Develoment, 19*(3), 627-641.
Jackson, Frank, Pettit, Philip, & Smith, Michael. (2000). Ethical Particularism and Patterns. In Brad Hooker & Margaret Little (Eds.), *Moral Particularism* (pp. 79-99). New York: Oxford University Press.
Jones, Karen. (2003). Emotions, Weakness of Will, and the Normative Conception of Agency. In Anthony Hatzimoysis (Ed.), *Philosophy and the Emotions* (pp. 181-200): Cambridge University Press.
Jones, Karen. (2006). Metaethics and Emotions Research: A Response to Prinz. *Philosophical Explorations, 9*(1), 45-53.
Joyce, Richard. (2006). *The Evolution of Morality*. Cambridge, MA: MIT Press.
Kahneman, Daniel. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist, 58*(9), 697-720.
Kahneman, Daniel, & Frederick, Shane. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In Thomas Gilovich, Dale Griffin & Daniel Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 49-81). New York: Cambridge University Press.
Kamm, Frances. (2009). Neuroscience and Moral Reasoning: A Note on Recent Research. *Philosophy and Public Affairs, 37*(4), 330-345.
Kant, Immanuel. (1785/1996). Groundwork of the Metaphysics of Morals. In Mary J. Gregor & Allen Wood (Eds.), *The Cambriidge Edition of the Works of Immanuel Kant: Practical Philosophy*: Cambridge University Press.
Kennett, Jeanette. (2006). Do Psychopaths Really Threaten Moral Rationalism? *Philosophical Explorations, 9*, 69-82.
Kennett, Jeanette, & Fine, Cordelia. (2009). Will the Real Moral Judgment Please Stand Up? The Implications of the Social Intuitionist Models of Cognition for

Meta-ethics and Moral Psychology. *Ethical Theory and Moral Practice, 12*(1), 77-96.
Kiehl, Kent. (2008). Without Morals: The Cognitive Neuroscience of Criminal Psychopaths. In Walter Sinnott-Armstrong (Ed.), *Moral Psychology Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (pp. 119-150). Cambridge, MA: MIT Press.
Koehler, Derek, Brenner, Lyle, & Griffin, Dale. (2002). The Calibration of Expert Judgment: Heuristics and Biases Beyond the Laboratory. In Thomas Gilovich, Dale Griffin & Daniel Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 686-715). New York: Cambridge University Press.
Kohlberg, L. (1984). *The Psychology of Moral Development: The Nature and Validity of Moral Stages*. New York: Harper & Row.
Korsgaard, Christine M. (1986). Skepticism about Practical Reason. *The Journal of Philosophy, 83*(1), 311-344.
Korsgaard, Christine M. (1996). *The Sources of Normativity*. New York: Cambridge University Press.
Krauthammer, Charles. (2009, May 1). Torture? No. Except... Op-Ed, *New York Times*. Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2009/04/30/AR2009043003108.html
Kunda, Ziva. (1990). The Case for Motivated Reasoning. *Psychological Bulletin, 108*(3), 480-498.
Lance, Mark, & Little, Margaret. (2004). Defeasability and the Normative Grasp of Content. *Erkenntnis, 61*, 435-455.
Lance, Mark, & Little, Margaret. (2005). Particularism and Antitheory. In David Copp (Ed.), *Oxford Handbook of Ethical Theory* (pp. 567-594). New York: Oxford University Press.
Lapsley, Daniel. (1996). *Moral Psychology*. Boulder, CO: Westview Press.
Lawrence, Gavin. (1995). The Rationality of Morality. In Rosalind Hursthouse, Gavin Lawrence & Warren Quinn (Eds.), *Virtues and Reasons: Philippa Foot and Moral Theory* (pp. 89-147). New York: Oxford University Press.
Little, Margaret. (2000). Moral Generalities Revisited. In Brad Hooker & Margaret Little (Eds.), *Moral Particularism* (pp. 276-304). New York: Oxford University Press.
Luntley, Michael. (2005). The Role of Judgment. *Philosophical Explorations, 8*(3), 281-295.
Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin Books.
Marr, David. (1982/2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.
McDowell, John. (1988a). Non-Cognitivism and Rule-Following. In *Mind, Value, and Reality* (pp. 198-219): Harvard University Press.
McDowell, John. (1988b). Two Sorts of Naturalism. In *Mind, Value, and Reality* (pp. 167-197): Harvard University Press.

McDowell, John. (1995). Might There Be External Reasons? In J.E.J. Altham & Ross Harrison (Eds.), *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams* (pp. 68-85). New York: Cambridge University Press.

McGuire, Jonathan, Langdon, Robyn, Coltheart, Max, & Mackenzie, Catriona. (2009). A Reanalysis of the Personal/Impersonal Distinction in Moral Psychology Research. *Journal of Experimental Social Psychology, 45*, 577-580.

McKeever, Sean, & Ridge, Michael. (2006). *Principled Ethics: Generalism as a Regulative Ideal*. New York: Oxford University Press.

Mikhail, John. (2007). Universal Moral Grammar: Theory, Evidence and the Future. *Trends in Cognitive Sciences, 11*(4), 143-152.

Mikhail, John. (2009). Moral Grammar and Intuitive Jurisprudence. *Psychology of Learning and Motivation, 50*, 27-100.

Mikhail, John. (2010). *Moral Grammar and Intuitive Jurisprudence: Theory, Evidence, and Future Research.* Paper presented at the Cognitive Science and Morality Workshop, Georgetown University.

Moll, Jorge, de Oliviera-Souza, Ricardo, Eslinger, Paul J., Bramati, Ivanei E., Mourao-Miranda, Janaina, Andreiuolo, Pedro Angelo, & Pessoa, Luiz. (2002). The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions. *The Journal of Neuroscience, 22*(7), 2730-2736.

Moll, Jorge, Eslinger, Paul J., & de Oliviera-Souza, Ricardo. (2001). Frontopolar and Anterior Temporal Cortex Activation in a Moral Judgment Task. *Arq Neuropsiquiatr, 59*(3-B), 657-664.

Muller, Jorgen L., Sommer, Monika, Wagner, Verena, Lange, Kirsten, Taschler, Heidrun, Roder, Christian H., Schuirerer, Gerhardt, Klein, Helmfried E., Hajak, Goran. (2003). Abnormalities in Emotion Processing within Cortical and Subcortical Regions in Criminal Psychopaths: Evidence from a functional Magnetic Resonance Imaging Study Using Pictures with Emotional Content. *Biological Psychiatry, 54*(2), 152-162.

Nagel, Thomas. (1979). *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.

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

Narvaez, Darcia. (2008). The Social Intuitionist Model: Some Counter-Intuitions. In Walter Sinnott-Armstrong (Ed.), *Moral Pscyhology Vol. 2: The Cognitive Science of Morality: Intuition and Diversity* (pp. 233-240). Cambridge, MA: MIT Press.

Nichols, Shaun. (2002). How Psychopaths Threaten Moral Rationalism: Is It Irrational to Be Amoral*? *The Monist, 85*(2), 285-304.

Nichols, Shaun. (2004). *Sentimental Rules*. New York: Oxford University Press.

Nickerson, Raymond S. (1998). Confirmation Bias: A Ubiquitious Phenomenon in Many Guises. *Review of General Psychology, 2*(2), 175-220.

Nisbett, Richard E., & Wilson, Timothy D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processs. *Psychological Review, 84*(3), 231-259.

O'Neill, P., & Petrinovich, L. (1998). A Preliminary Cross-Cultural Study of Moral Intuitions. *Evolution and Human Behavior, 19*(6), 349-367.
Oakley, Justin. (1992). *Morality and the Emotions*. New York: Routledge.
Memorandum for John A. Rizzo, Senior Deputy General Counsel, Cental Intelligence Agency: Re: Application of United States Obligations Under Article 16 of the Convention Against Torture to Certain Techniques that May Be Used in the Interrogation of High Value al Qaeda Detainees (May 30, 2005).
Payne, B. Keith. (2005). Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping. *Journal of Personality and Social Psychology, 89*(4), 488-503.
Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An Empirical Study of Moral Intuitions: Toward an Evolutionary Ethics. *Journal of Personality and Social Psychology, 64*(3), 467-478.
Petrinovich, Lewis, & O'Neill, Patricia. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology, 17*(3), 145-171.
Piaget, Jean. (1932/1965). *The Moral Judgment of the Child* (Marjorie Gabain, Trans.). New York: The Free Press.
Pizarro, David A., & Bloom, Paul. (2003). The Intelligence of Moral Intuitions: Comment on Haidt (2001). *Psychological Review, 116*(1), 193-196.
Prinz, Jesse. (2007). *The Emotional Construction of Morals*. New York: Oxford University Press.
Pylyshyn, Zenon W. (1981). The Imagery Debate: Analog Media Versus Tacit Knowledge. In Ned Block (Ed.), *Imagery*. Cambridge, MA: MIT Press.
Rachels, James. (1993). Subjectivism. In Peter Singer (Ed.), *A Companion to Ethics*. Oxford: Blackwell.
Railton, Peter. (2006). Normative Guidance. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics, Volume 1* (pp. 3-34). New York: Oxford University Press.
Rawls, John. (1951). Outline of a Decision Procedure for Ethics. *Philosophical Review, 60*(2), 177-197.
Rawls, John. (1999). *A Theory of Justice, Revised Edition* (Revised Edition ed.). Cambridge, MA: Harvard University Press.
Roskies, Adina. (2003). Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy". *Philosophical Psychology, 16*(1), 51 - 66.
Roskies, Adina. (2006). Patients With Ventromedial Frontal Damage Have Moral Beliefs. *Philosophical Psychology, 19*(5), 617 - 627.
Ross, W. David. (1930). *The Right and the Good*. New York: Oxford University Press.
Rozin, Paul, Lowery, Laura, Imada, Sumio, & Haidt, Jonathan. (1999). The CAD Triad Hypothesis: A Mapping between Three Moral Emotions (Contempt, Anger, Disgust) and Three Moral Codes (Community, Autonomy, Divinity). *Journal of Personality and Social Psychology, 76*(4), 574-586.
Ryle, Gilbert. (1949). *The Concept of Mind*. Chicago: The University of Chicago Press.
Sadler, Brook J. (2003). The Possibility of Amoralism: A Defence Against Internalism. *Philosophy, 78*, 63-78.

Saltzstein, Herbert D., & Kasachkoff, Tziporah. (2004). Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique. *Review of General Psychiatry, 8*(4), 273-282.
Saver, J.L., & Damasio, A. R. (1991). Preseved Access and Processing of Social Knowledge in Patient with Acquired Sociopathy Due to Ventromedial Frontal Damage. *Neuropyschologia, 29*(12), 1241-1249.
Scanlon, T. M. (2000). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as Embodied Moral Judgment. *Personality and Social Psychology Bulletin, 34*(8), 1096.
Schwarz, Norbert. (2002). Feelings as Information: Moods Influence Judgments and Processing Strategies. In Thomas Gilovich, Dale Griffin & Daniel Kahneman (Eds.), *Heuristics and Biases* (pp. 534-547): Cambridge University Press.
Shiffrin, Seana Valentine. (1999). Moral Overridingness and Moral Subjectivism. *Ethics, 109*(4), 772-794.
Singer, Peter. (1974). All Animals Are Equal. *Philosophical Exchange, 1*, 103-116.
Singer, Peter. (2005). Ethics and Intuitions. *The Journal of Ethics, 9*, 331-352.
Sinnott-Armstrong, Walter. (1999). Some Varieities of Paticularism. *Metaphilosophy, 30*(1/2), 1-12.
Sinnott-Armstrong, Walter, Young, Liane, & Cushman, Fiery. (2010). Moral Intuitions as Heuristics. *Unpublished manuscript*.
Sloman, Steven A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin, 119*(1), 3- 22.
Sloman, Steven A. (2002). Two Systems of Reasoning. In Thomas Gilovich, Dale Griffin & Daniel Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 379-396). New York: Cambridge University Press.
Smith, Eliot R., & DeCoster, Jamie. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality and Social Psychology Review, 4*(2), 108-131. doi: 10.1207/s15327957pspr0402_01
Smith, Michael. (1994). *The Moral Problem*. Malden, MA: Blackwell.
Spelke, Elizabeth S. (2000). Core Knowledge. *American Psychologist, 55*, 1233-1243.
Sripada, Chandra Sekhar, & Stich, Stephen. (2006). A Framework for the Psychology of Norms. In Peter Carruthers, Stephen Laurence & Stephen Stich (Eds.), *The Innate Mind Volume 2: Culture and Cognition*. New York: Oxford University Press.
Stanovich, Keith E. (1999). *Who Is Rational? Studies of Individual Differences in Reasoning*. New York: Psychology Press.
Stanovich, Keith E. (2009). *What Intelligence Tests Miss: The Psychology of Rational Thought*. New Haven, CT: Yale University Press.
Stanovich, Keith E., & West, Richard F. (1988). Individual Differences in Rational Thought. *Journal of Experimental Psychology, 127*(2), 161-188.
Stanovich, Keith E., & West, Richard F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioral and Brain Sciences, 23*, 645-726.

Sterelny, Kim. (2008). The Fate of the Third Chimpanzee. *Jean Nicod Lectures*.
Stevenson, C. L. (1937). The Emotive Meaning of Ethical Terms. *Mind, 46*(181), 14-31.
Stich, Stephen. (2006). Is Morality an Elegant Machine or a Kludge? *Journal of Cognition and Culture, 6*(1-2), 181-189.
Stocker, Michael. (1979). Desiring the Bad: An Essay in Moral Psychology. *Journal of Philosophy, 76*(12), 738-753.
Strack, Fritz, & Deutsch, Roland. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review, 8*(3), 220-247.
Strawson, Peter. (1962). Freedom and Resentment. *Proceedings of the British Academy, 48*, 1-25.
Sunstein, Cass R. (2005). Moral Heuristics. *Behavioral and Brain Sciences, 28*, 531-573.
Svavarsdottir, Sigrun. (1999). Moral Cognitivism and Motivation. *Philosophical Review, 108* (2), 161-219.
Thomson, Judith Jarvis. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist, 59*, 204-217.
Tiberius, Valerie. (2008). *The Reflective Life: Living Wisely Within Our Limits*. New York: Oxford University Press.
Turiel, Elliot. (1983). *The Development of Social Knowledge: Morality & Convention*. New York: Cambridge University Press.
Tversky, Amos, & Kahneman, Daniel. (1983). Extensional vs. Intuitional Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychology Review, 90*, 293-315.
Valdesolo, Piercarlo, & DeSteno, David. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science, 17*(6), 476-477.
Vaughn, Lewis. (2008). *Doing Ethics: Moral Reasoning and Contemporary Issues*. New York: W.W. Norton.
Wallace, James D. (2008). *Norms and Practices*. Ithaca, NY: Cornell University Press.
Wheatley, Thalia, & Haidt, Jonathan. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science, 16*(10), 780-784.
Wielenberg, Erik J. (2010). On the Evolutionary Debunking of Morality. *Ethics*, 441-446.
Wiggins, David. (1987). A Sensible Subjectivism? In *Needs, Values, Truth: Essays in the Philosophy of Value* (pp. 185-214). New York: Oxford University Press.
Williams, Bernard. (1973a). A Critique of Utilitarianism. In J.J.C. Smart & Bernard Williams (Eds.), *Utilitarianism: For and Against* (pp. 77-150). New York: Cambridge University Press.
Williams, Bernard. (1973b). *Problems of the Self*. New York: Cambridge University Press.
Williams, Bernard. (1981). Internal and External Reasons. In *Moral Luck* (pp. 101-113). New York: Cambridge University Press.
Williams, Bernard. (1986). *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Wilson, Timothy D., Lindsey, Samuel, & Schooler, Tonya Y. (2000). A Model of Dual Attitudes. *Psychological Review, 107*(1), 101-126.

Winkielman, Piotr, Zanjonc, Robert B., & Schwarz, Norbert. (1997). Subliminal Affective Priming Resists Attributional Interventions. *Cognition and Emotion, 11*(4), 433-465.

Young, Liane, Camprodon, Joan Albert, Hauser, Marc D., Pascual-Leone, Alvaro, & Saxe, Rebecca. (2010). Disruption of the Right Tempoparietal Junction with Transcranial Magnetic Simulation Reduces the Role of Beliefs in Moral Judgments. *Proceedings of National Academy of Sciences, 107*(15), 6753-6758.

Zagzebski, Linda. (2003). Emotion and Moral Judgment. *Philosophy and Phenomenological Research, 66*(1), 104-124.