Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Jared K Rothstein November 3, 2011

The Affective Foundations of Moral Cognition and Justification: A Naturalistic Account

By

Jared K Rothstein
Doctor of Philosophy

Philosophy

Dr. Robert N. McCauley
Advisor

Dr. Mark Risjord
Committee Member

Dr. Frans de Waal
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

The Affective Foundations of Moral Cognition and Justification: A Naturalistic Account

By

Jared K Rothstein
M.A., Emory University, 2007

Advisor: Robert N. McCauley, Ph.D.

An abstract of
A dissertation submitted to the faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Philosophy
2011

Abstract

The Affective Foundations of Moral Cognition and Justification: A Naturalistic Account

By Jared K Rothstein

This interdisciplinary project focuses on the vital role of emotion in moral cognition and the ramifications for a psychologically realistic approach to normative ethical reasoning. Convergent evidence from a variety of scientific fields, including psychology, neuroscience, and primatology, indicates that affect directs our intuitive judgments, grounds our empathic capacities, orients our moral reasoning, and motivationally binds us to our assessments. These descriptive findings shed light on several prominent metaethical issues, including the simulation/'theory theory' debate, the weakness of will question, and the realism/antirealism controversy. With regard to the simulation/'theory theory' debate concerning the neuropsychological underpinnings of our Theory of Mind (ToM) capacities, it appears that both sides are partially correct. As reflected by the distinctive empathic impairments characteristic of psychopathy and autism and supporting neurological evidence, Affective ToM and Cognitive ToM rely on unique underlying mechanisms, with the former incorporating more simulation-based processing and the latter involving more theory-based operations. In Chapter 3, it is argued that weakness of will occurs less frequently in the case of intuitive judgments, as opposed to assessments based on conscious moral reasoning, since intuitive judgments are typically linked to relatively more intense emotion and thus carry greater motivational force. Furthermore, I contend in Chapter 4 that a Darwinian genealogy of our ethical sensibility poses a serious epistemological challenge to traditional versions of moral realism, a view that there are 'independent' ethical truths that apply irrespective of our subjective feelings. The practical implications of this evolutionary debunking are limited, however, because our tendency to impute greater practical authority to ethical norms is emotionally-rooted, persisting in the absence of a belief in moral realism. Finally, in my last chapter, I endorse and expand John Rawls' method of wide reflective equilibrium as a psychologically realistic approach to ethical justification that accords with the empirical evidence regarding the affective foundations of moral judgment and motivation. To my awareness, this enhanced version of Rawls' method is the first developed normative reasoning procedure of its kind within the sentimentalist tradition—a justificatory approach that attributes normative weight to our moral feelings without, however, automatically justifying them.

The Affective Foundations of Moral Cognition and Justification: A Naturalistic Account

By

Jared K Rothstein
M.A., Emory University, 2007

Advisor: Robert N. McCauley, Ph.D.

A dissertation submitted to the faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Philosophy
2011

Table of Contents

1. The Social Intuitionist Model of Moral Judgment

"Reason is, and ought only to be the slave of the passions…"

--David Hume (1793/1964)

Embraced by sympathizers and reviled by rationalist critics, Hume's bold pronouncement—an enduring testament of the sentimentalist position in ethics—has been a lightning rod of controversy over the years. Even among sentimentalists, there is no consensus regarding the implications. As I construe Hume's seminal statement, it is fundamentally about the vital role of emotion in moral cognition and the ramifications for a psychologically realistic approach to normative ethical reasoning. Indeed, there are two main facets to Hume's argument, a descriptive and a normative dimension; the latter of which has received substantially less attention from contemporary sentimentalists despite its great importance. The descriptive claim is relatively straightforward. According to Hume's psychological theory, emotion drives everyday moral judgment. Reason may help us to frame or categorize ethical situations, but our feelings ultimately determine how these situations are assessed. As will be outlined in this project, contemporary research from a number of scientific fields, including neuroscience, cognitive and developmental psychology, and primatology, supports Hume's basic contention that emotion plays a central role in moral cognition. It will be shown that affect directs our intuitive judgments, grounds our empathic capacities, orients our moral reasoning, and motivationally binds us to our assessments.

The other dimension of Hume's famous quote remains much more controversial. In line with his psychological observations, Hume made the *normative* claim that reason *ought* to be the slave of the passions. The implications of this statement have been the

subject of much debate within the sentimentalist tradition. Contemporary, empirically-minded sentimentalists, however, have generally steered clear of this thorny normative question, focusing instead on the upshot of scientific research for descriptive ethics. For Hume, however, the descriptive and the normative are inextricably connected. His normative claim follows directly from his empirical observations. In my view, Hume's claim that reason *ought* to be the slave of the passions is fundamentally an insight about what a psychologically realistic approach to moral justification must look like. Hume intimates that the normative ethical prescriptions that philosophers reflectively endorse must accord with our gut feelings, or else they will be of little practical value, since we cannot be expected to accept and act upon prescriptions that run counter to our emotional commitments. Hume offered his bold normative claim as a corrective to the longstanding rationalist view in ethics that our moral feelings are without justificatory force, a position endorsed by his rival, Kant (1785/1964), and still widely prevalent among contemporary ethicists today.

It seems, however, that Hume was right, based on the empirical evidence regarding ethical judgment and motivation reviewed in Chapters 1-4. That is, in order to be psychologically realistic, a justificatory reasoning procedure in ethics must factor in our emotional commitments and be able to generate a set of moral prescriptions that largely cohere with our gut feelings. In Chapter 5, I endorse and enhance John Rawls' (1971/1999) method of wide reflective equilibrium as a naturalistic approach to moral justification that attributes normative weight to our moral feelings, without automatically justifying all of them. To my awareness, this is the first *developed* normative reasoning model of its kind within the sentimentalist tradition—a justificatory approach that

successfully bridges the descriptive and normative dimensions of ethical theorizing in accord with Hume's original purpose. Before this normative reasoning procedure can be outlined, however, we must first turn our attention to the descriptive findings that lay the groundwork. The driving force of gut feeling in intuitive ethical judgment will be the focus of this chapter.

I. Trolley Problems and Moral Dumbfounding

In recent years, a growing number of researchers have addressed the role of moral intuition in ethical judgment and reasoning offering exciting new insight regarding the underlying mechanisms. The findings indicate that many of our everyday ethical assessments are snap judgments, based on hot, emotionally-laden intuitions, rather than conscious moral reasoning. Research has consistently revealed a large gap between the ethical assessments we issue and our ability to provide consistent, rational justification for them. "Moral dumbfounding," a phenomenon in which subjects struggle to justify the moral convictions they report, has been commonly reported in studies of this type.

Perhaps the most widely cited research in the area of intuitive ethical judgment is the so-called "trolley problem" studies. Various versions have been conducted by several researchers (see Hauser et al., 2008, pg. 127, for a comprehensive listing), revealing recurrent patterns of judgment and reasoning across diverse populations of subjects. In these studies, subjects are asked to assess the moral permissibility of various versions of a similar scenario--a train is heading toward 5 doomed individuals on a track and an onlooker has an opportunity to save them by sacrificing an innocent bystander. In each version, the onlooker utilizes a different means of sacrifice. There are two primary

scenarios utilized in this research. In the Standard condition, the onlooker *flips a switch*, diverting the train onto a side track where the bystander is located. The Footbridge condition, in contrast, requires that the onlooker *push* the bystander in front of the oncoming train. In general, the vast majority of subjects approve of the action taken in the Standard condition, while disapproving of the Footbridge scenario. Joshua Greene (2008) explains this pattern of divergent judgment by proposing that we have an evolved aversion to acts involving direct, personal harm—an emotional response that is triggered by the Footbridge scenario, but not the Standard condition. As discussed in Section III below, Greene provides fMRI evidence in support of this hypothesis.

Other researchers, such as Marc Hauser (2006), propose an alternative explanation for why we tend to judge in favor of the Standard condition while rejecting the Footbridge scenario. According to Hauser's hypothesis, these intuitive judgments are (unconsciously) based on a principle of "double effect"---holding that otherwise prohibited acts may be permissible if the good effects outweigh the bad, and the harm they cause is unintentional. The Standard and Footbridge scenarios both satisfy the first, utilitarian criterion (i.e., one innocent bystander is killed to save five lives). Only the Standard scenario, however, fulfills the second requirement involving unintentional harm. In the Footbridge condition, the harm caused to the innocent bystander (i.e., pushing him in front of the train) is an *intended* means of saving the five lives, whereas in the Standard condition this sacrifice is an *unintended* consequence of redirecting the train onto the side track. In order to test this "double effect" hypothesis, Hauser et al. (2008) tested subjects on two additional, "loop scenarios," neither of which involve personal harm, thus controlling for this variable. In the Direct Loop condition, the onlooker flips a switch to

divert a train onto a side track that loops back onto the main track in front of where the five innocent individuals are standing. Subjects are told that standing on this side track is an innocent bystander who will be hit, slowing the train enough to afford the five potential victims on the main track time to escape. Hence, as in the Footbridge scenario, the Direct Loop condition involves a violation of the principle of "double effect," since the harm caused to the innocent bystander on the side track is an intended means of saving the five lives (i.e., hitting the innocent bystander in order to slow the train's momentum). The Indirect Loop scenario is identical to the Direct Loop condition except that in this scenario the innocent bystander standing on the loop is positioned in front of a heavy object. Thus, just like in the Standard condition, the Indirect Loop scenario involves killing the innocent bystander as *a foreseen, but unintended consequence* of saving the five lives (i.e., the direct means is redirecting the train into the heavy object).

In the broadest study to date (Hauser et al., 2008) of judgments regarding these four trolley problem scenarios, encompassing over 30 000 subjects from 120 countries, 89% of subjects approved of the Standard condition while judging the Footbridge scenario to be morally impermissible. A smaller majority judged in favor of the two loop conditions: 72% for the Indirect Loop case and 55% for the Direct Loop scenario. What are the implications of these findings for the debate between Greene and Hauser regarding the underlying psychological explanation for these patterns of intuitive judgment? It appears that Greene has the upper hand. His theory that our divergent judgments regarding the two primary trolley problem scenarios, the Footbridge and the Standard condition, stems from a difference in the emotional response elicited by each is independently supported by his fMRI research, and there is a high degree of consensus

among subjects regarding these two scenarios in a pattern that accords with Greene's explanation. Hauser's "double effect" hypothesis, on the other hand, lacks independent empirical support, and the findings regarding the two Loop scenarios do not provide compelling evidence. Although more subjects judged in favor of the Indirect Loop scenario than Direct Loop condition, as predicted by Hauser's hypothesis, there was relatively less consensus regarding these two Loop conditions. More importantly, 55% of subjects judged in favor of the Direct Loop scenario, a condition which violates the principle of "double effect." As Greene (2008) emphasizes, "[more than] half the subjects *do the opposite of what [Hauser's] theory predicts*" (pg. 112). At the same time, Greene acknowledges that his 'personal harm' explanation requires further refinement and needs to be supplemented by a richer account of how emotion shapes a wider variety our intuitive ethical assessments. He writes, "my current opinion is that both [my explanation and Hauser's] are incomplete and descriptively inadequate" (pg. 106).

Setting this issue to the side, one of the most striking aspects of this trolley problem research program is the discovery of recurrent patterns of intuitive judgment, especially with regard to the Standard and Footbridge scenarios, across subjects from a variety of cultural and ethnic backgrounds. As noted above, in Hauser et al.'s (2008) study, over 30 000 subjects hailing from 120 different countries were tested utilizing web-based technology. Analysis of the initial data set, which included 5000 subjects of distinctive national, religious, and ethnic affiliations, revealed no significant difference in judgment patterns regarding the Standard and Footbridge scenarios across these demographic sub-sets (Hauser et al., 2008, pg. 130). Providing further evidence that these intuitive biases are widely shared, Hauser et al. report that Cristopher Marlowe has

uncovered similar assessment trends among the Hadza, a small group of hunter-gatherers living in a remote area of Tanzania. These trolley problem findings demonstrating recurrent patterns of intuitive ethical judgment across cultures are consistent with an *innate biases* view of intuitive ethical judgment, championed by Greene (2008) and Jonathon Haidt (2001). As outlined in Section IV below, according to this view, many of our intuitive assessments are rooted in evolved affective predispositions, leading to patterns of cross-cultural convergence.

In addition to revealing recurrent patterns of intuitive assessment, these trolley problem studies also demonstrate our general inability to provide adequate justification for our gut intuitions. In the web-based study by Hauser et al. (2008), after being presented with both the Standard and Footbridge scenarios and offering their immediate judgments, subjects were asked to describe the rationale for their verdicts. Applying a relatively charitable standard, Hauser et al. defined a sufficient justification as "one that correctly identified any factual difference between the two scenarios and claimed the difference to be the basis of moral judgment" (pg. 130). Experimenters did not assess the normative weight of the factual distinctions cited by subjects. Nonetheless, only 30% of subjects provided 'sufficient justification.' Researchers found no correlation between subjects' age, gender or religious affiliation and the likelihood of providing adequate reasons. Those with a background in moral philosophy, however, were more likely to provide adequate justification (pg. 131). Subjects performed even worse when asked to explain their judgments regarding both the Direct Loop and Indirect Loop scenarios, with only 13% providing 'sufficient justification.' Based on these findings, Hauser et al.

conclude: "there is a disassociation between judgment and justification, suggesting that intuition as opposed to principled reasoning guides judgment" (pg. 133).

Consistent with these findings, in a series of studies conducted by Jonathon Haidt and colleagues (2001, 2000, 1993), researchers found that, when asked to assess hypothetical scenarios, subjects generally make snap judgments and remain committed, even when struggling to justify these gut assessments. For instance, in an experiment conducted by Haidt, Koller, and Dias (1993), 360 subjects, having diverse national, age and class affiliations, were presented with scenarios describing harmless taboo violations (e.g., a family eats a pet dog after it was killed in an accident; an old national flag is used to scrub a toilet; etc.). Subjects were asked by investigators if these scenarios depicted moral violations or not, and why. Interestingly, researchers found that individuals of high social class generally did not view these taboos as moral in character, while subjects having a lower socio-economic status did—indicating that socioeconomic factors can influence at least some of our intuitive assessments. Regardless of the verdict rendered, however, most subjects appeared to rely on gut intuition in issuing their judgments. In response to initial questioning, many struggled to justify their convictions. With further questioning from investigators, several subjects admitted to being dumbfounded, unable to explain why they reached their verdict; which generally did not lead to any revision of their original assessment, however. Haidt and Hersh (2001) uncovered the same broad pattern of moral dumbfounding in a replication study investigating the judgments of political liberals and conservatives regarding various forms of sexual behavior involving masturbation, homosexuality, and incest. Reporting the findings from a similar study he conducted with Bjorklund and Murphy (2000), Haidt (2008) describes the attempts of

subjects to rationalize their intuitive judgments in the following way: "[a] very quick judgment was followed by a search for supporting reasons only; when these reasons were stripped away by the experimenter, few subjects changed their mind, even though many confessed that they could not explain the reasons for their decisions" (pg. 198).

The findings of Haidt and colleagues provide further support for the basic conclusion drawn by Hauser et al (2008). That is, it appears that many of our everyday moral judgments are based on immediate intuition; and for this psychological class of assessment, ethical reasoning is more often a biased search for justification rather than an open-ended deliberation. Even when we struggle to find supporting reasons for our intuitive biases, we, nonetheless, typically remain committed to them--a tendency reflected by the phenomenon of "moral dumbfounding."

II. The Moral Grammar Model

The studies discussed above indicate that many of our everyday moral judgments are driven by intuition rather than conscious moral reasoning. Accordingly, there has recently been a strong push within the field of moral psychology to better understand the underlying mechanisms of intuitive ethical assessment. Today, there are two primary, competing accounts: Hauser's (2006) "Moral Grammar" model and Haidt's (2001) "Social Intuitionist" account. I will argue that the latter has the most empirical support, given the compelling evidence that emotion plays a key role in intuitive ethical judgment.

. The chief alternative to Haidt's sentimentalist theory is a view endorsed by Marc Hauser (2006). While agreeing with Haidt that everyday ethical judgments are issued relatively automatically, Hauser proposes a unique account of the underlying

mechanisms—suggesting that affect plays no direct role. Drawing an analogy to Chomsky's nativist theory of "universal grammar," Hauser argues that intuitive ethical assessments are based on the unconscious processing of universally-shared moral principles. According to this view, just as we are born with a set of universal linguistic rules or principles that constrain the possible forms of human language, we also come equipped with an innate "moral grammar," operating in a similar way. Hauser writes, "[we] are endowed with a *moral faculty*—a capacity that enables each individual to unconsciously and automatically evaluate a limitless variety of actions in terms of principles that dictate what is permissible, obligatory, or forbidden" (pg. 36). Hauser proposes (pg. 44) that these inborn principles—which were adaptive for our ancestors--serve as universally-shared general rules (e.g., murder is wrong), with culture determining the exceptions (e.g., killing cheating spouses is obligatory in some cultures). He writes, "[we have] a suite of principles and parameters for building moral systems. These principles lack specific content….What gives these principles content is the local culture" (pg. 298). In claiming that these general principles or rules lack "specific content," Hauser is acknowledging the important role of cultural specification in shaping their "parameters." Again, the parameters to which Hauser is referring are local exceptions to the universally-shared rules we possess. For example, we might all endorse the broad principle that 'lying is wrong,' but what constitutes a genuine case of lying may vary from culture to culture. Is there a difference between explicitly lying and merely failing to disclose the truth? Is it permissible to tell a 'white lie' to protect another's feelings? Hauser acknowledges that local culture can influence the relatively more

subtle discriminations we make regarding how to apply general rules to concrete cases, while insisting, nonetheless, that these broad principles are universally-shared.

Hauser et al. (2008) further clarify how local culture can prune our innate moral grammar. According to Hauser et al.'s model of intuitive ethical judgment, there are two primary stages of processing. The first, *action analysis* phase deals with the perception and categorization of novel situations. Once we intuitively categorize a situation (e.g., 'X is a case of lying'), this initiates the *assessment* phase in which we unconsciously match the categorized situation to a relevant moral rule (e.g., 'lying is wrong') and judge accordingly ('X is wrong'). Although Hauser et al. do not explicitly draw the connection, it is clear that the cultural pruning to which they refer occurs primarily at the *action analysis* phase of judgment. Again, as characterized by Hauser above, local culture determines the range of actions that properly fall within a given category (e.g., when an action should be perceived as a case of 'lying'), which allows for cultural exceptions to the general principles constitutive of our innate moral grammar. Indeed, Hauser et al. criticize competing theories of intuitive ethical judgment, such as the one provided by Haidt (2001), for failing to incorporate an adequate account of this important *action analysis* phase. Hauser et al. write,

> it will not do to merely assign the role of moral judgment to reason, emotion or both. We must describe the computations underlying the judgments we produce….Minimally each of the other models must recognize an appraisal system that computes the causal-intentional structure of an agent's actions and the consequences that follow (pg. 117).

Hauser et al. argue that any adequate theory of intuitive moral judgment must explain how the mind initially perceives and categorizes moral situations. For their part, Hauser et al. propose that ethical actions are unconsciously perceived along several key dimensions—e.g., who is the agent (adult, child, adolescent, etc.), who is the recipient, what is the agent's relationship to patient, what is the agent's intention, what are the consequences of the agent's action, etc. Moral situations are categorized based on these variables, which then triggers the unconscious application of a corresponding ethical principle.

According to this cold-processing view, intuitive ethical judgment (from the *action analysis* to the *assessment* phase) occurs in the absence of affective influence. On this account, our automatic assessments elicit emotional responses, not the other way around. Hauser et al. (2008) write,

> the operative principles of the moral faculty do all the heavy lifting, generating a moral verdict that may or may not generate an emotion or a process of rational and principled deliberation… Emotion and deliberate reasoning are not causally related to our initial moral judgment, but, rather, are caused by the judgment (pg. 117-121).

Distinguishing between moral competence and behavior, Hauser et al. theorize that affect only impacts the latter, playing a motivational role. That is, once an intuitive judgment is issued (moral competence), this typically triggers an emotional response, which then motivates action in accord with the assessment. Accordingly, Hauser (2006) speculates that psychopaths are morally competent but lack the requisite emotional repertoire to follow through on their judgments. As emphasized in the next chapter, psychopathy

research contradicts Hauser's contention that moral competency is spared in this population. It appears the emotional deficits characteristic of this disorder do, in fact, lead to aberrant assessments. Indeed, the next section focuses on research indicating that emotion typically drives intuitive ethical judgment, *contra* Hauser et al.'s contention.

III. Intuitive Ethical Judgment and Emotion

Proponents of the Moral Grammar model theorize that intuitive ethical judgment occurs in the absence of affective influence. This view lacks empirical support. As outlined below, convergent evidence from a variety of fields supports an alternative, sentimentalist account. Indeed, it appears that gut feeling often determines the snap assessments we issue.

Antonio Damasio (2000) presents research consistent with this view. Based on his studies of VM patients (i.e., individuals with damage to their ventromedial prefrontal cortex, an area of the brain linked to emotional processing), Damasio argues that affective tags or "somatic markers" play a necessary role in social and personal decision-making. Many of his patients, despite having an abstract understanding of basic moral principles and social convention and a capacity to assess risk, are unable to effectively put this knowledge into practice. In attempting to make decisions, they have trouble weighing different options, often spending an inordinate amount of time deciding what to do, while eventually making seemingly irrational choices (e.g., taking irresponsible risks, ignoring obligations, etc.). Damasio accounts for this deficit in terms of his "somatic marker hypothesis." According to this theory, in normal subjects, decision-making is guided by intuitive affective responses—somatic markers or tags—connected to various

options. These markers circumscribe the range of viable alternatives, while providing the impetus to choose one over another. For instance, in considering what I should do today, some options feel 'out of the question' (e.g., taking a long road trip), while I am compelled toward others (e.g., continue working on this paper). Damasio theorizes that VM patients lack these somatic markers, and this explains their decision-making deficits. Damasio's theory as it pertains to ethical cognition is discussed in greater depth in Chapter 3.

In accord with Damasio's general insight that affect plays a key role in moral judgment, researchers have demonstrated that emotional manipulation can alter our intuitive assessments. In an experiment conducted by Haidt and Wheatley (2005), highly hypnotizable subjects were given a posthypnotic suggestion (of which they were unaware) to experience disgust whenever they read a trigger word. Half of the subjects were primed for the word 'take,' while the other half were primed for 'often.' Subjects were presented with six moral scenarios, containing one of the two trigger words. Haidt and Wheatley found that subjects' moral judgments were more severe for the disgust-inducing scenarios. In a replication study, Haidt and Wheatley included a seventh scenario in which no violation (i.e., neither conventional nor moral) was described. Remarkably, one third of the subjects judged the actions described as 'somewhat morally wrong,' apparently misattributing the aversive feelings elicited by the hidden trigger word. As reported by Haidt and Bjroklund (2008), the general finding that manipulating disgust reactions can make subjects' ethical judgments more severe has been replicated in two additional studies (Haidt and Bjorklund, unpublished; Shnal et al., 2007).

Shedding additional light on this phenomenon, Shaun Nichols (2004a) conducted a revealing study of disgust norms, a group of norms typically linked to intense affect. He found that subjects judged violations of disgust norms very similarly to moral violations along several key dimensions. As compared to violations of conventional social norms (e.g., putting your elbows on the table), which appear to be less emotionally-charged, violations of both moral and disgust norms tend to be judged as more serious, less permissible, and less authority-contingent. Nichols argues that these findings provide strong analogical evidence that moral assessments, like judgments of disgust, are rooted in intense affective responses; which explains why both types of normative violations are judged in a similar fashion. In support of this hypothesis, Nichols reports that psychopaths—a population known to have affective deficits—fail to draw the standard distinction between moral and conventional normative violations. Chapter 4 addresses the implications of this moral/conventional research in greater detail.

Additional findings offer more direct evidence that our intuitive ethical judgments are emotionally-driven. Koenigs et al. (2007) conducted a revealing study of VM patients—testing them on a variety of moral dilemmas, including the basic set of trolley problem scenarios. These subjects deviated significantly, statistically speaking, from normal populations in their judgments regarding scenarios that involved direct, personal harm, such as the Footbridge condition. As opposed to their normal counterparts, VM patients judged these 'personal harm' scenarios to be morally permissible, issuing a utilitarian-style assessment (i.e., 'it is permissible to sacrifice one life to save five, even if this requires inflicting direct, personal harm on the individual being sacrificed'), whereas normal subjects tend to deliver a deontological-style verdict (i.e., 'it is impermissible to

inflict direct, personal harm on an individual, even if this would save five lives'). In response to these findings, Hauser et al. (2008), proponents of the cold-processing, Moral Grammar model of intuitive ethical judgment, reluctantly acknowledge that "in this selective set of moral problems, emotions appear causally necessary. When the circuitry subserving social emotions is damaged, a hyper-utilitarian emerges" (pg. 138).

Greene (2008) presents fMRI findings that may help to explain Koenig et al.'s (2007) results. In a series of neuroimaging studies conducted by Greene and colleagues, researchers found that contemplation of 'personal harm' trolley scenarios, like the Footbridge—which typically elicit deontological-style, rather than utilitarian-style verdicts--is associated with relatively greater activity in 'emotional' brain regions, the posterior cingulate cortex, the medial prefrontal cortex, and the amygdala. By comparison, contemplation of 'impersonal' trolley dilemmas, such as the Standard condition, involves relatively greater activity in classically "cognitive" regions, the dorsolateral prefrontal cortex and inferior parietal lobe (pg. 44). Based on these findings, Greene speculates that different neurological mechanisms underlie deontological-style versus utilitarian-style judgments—and that the former class of assessments are more intuitive (i.e., relatively quicker, more automatic, and based less on conscious moral reasoning) than the latter class. He writes, "deontological judgments tend to be driven by emotional responses…. This is in contrast to [consequentialist judgments], which…arise from rather different psychological processes, ones that are more 'cognitive,' and more likely to involve genuine moral reasoning" (pg. 36). Greene contends that intuitive, deontological-style assessments are characteristically driven by more intense affective responses—emotional "alarm bells"—whereas utilitarian-style judgments are influenced

by more subtle affective cues, what Damasio identifies as "somatic markers." Accordingly, Greene claims that each psychological type of judgment is typically linked to a distinctive kind of moral 'reasoning." We tend to merely *rationalize* our intuitive assessments, in contrast to the more open-ended deliberation characteristic of utilitarian-style judgments. Greene argues, "the only way to reach a distinctively consequentialist judgment (i.e., one that doesn't coincide with a deontological judgment) is to actually go through the consequentialist, cost-benefit reasoning using one's 'cognitive' faculties, the ones based in the dorsolateral prefrontal cortex (pg. 65). According to this view, VM patients offer aberrant, utilitarian-style verdicts to moral scenarios involving 'personal harm' because their affective deficits short-circuit the emotional alarm bells characteristically elicited in normal subjects. This, in turn, allows VM patients to engage in utilitarian-style reasoning and judge accordingly; whereas normal subjects are driven by a strong gut feeling to issue a snap, deontological-style verdict.

In Chapter 3, I endorse the basic psychological distinction Greene draws between two classes of judgment, intuitive/alarm bell and reason-based/somatic marker judgments. That is, I agree with Greene that there are strong grounds for psychologically distinguishing intuitive judgments from reason-based ones according to the unique type of emotion--alarm bell versus somatic marker--characteristically linked to each. I reject, however, Greene's further suggestion the intuitive/alarm bell judgments will only result in deontological-style verdicts and that utilitarian-style verdicts may only be reached on the basis of conscious moral reasoning. In other words, *contra* Greene's suggestion, the psychological distinction between intuitive/alarm bell and reason-based/somatic marker judgments *does not* overlap with a functional distinction (i.e., based on the type of verdict

reached) between deontological-style and utilitarian-style assessments. For present purposes, however, the central point is that Greene's fMRI research provides compelling evidence that at least some types of intuitive judgment (e.g., those elicited in response to 'personal harm' scenarios, like the Footbridge) are driven by strong affective responses. This provides further support for a hot-processing account of intuitive ethical assessment.

Evolutionary considerations also recommend this view. In recent years, several theorists, de Waal (2006; 1996), Hauser (2006), L. Arnhart (1998), and S. Pinker (1997), just to name a few, have written about the evolutionary origins of human morality. What once seemed like a puzzle—how evolution could favor the development of creatures who exhibit altruistic behavior—now seems relatively clear, based on the theories of inclusive fitness and reciprocal altruism. As underscored by R. Dawkins (2006) in the *Selfish Gene*, a marker of evolutionary fitness is the survival and reproduction of an individual's genes. Hence, sacrificing for kin, who share much of our genetic make-up, may still be adaptive. Furthermore, as highlighted by R. Trivers (1971), who first developed the theory of reciprocal altruism, helping unrelated individuals can also be an adaptive strategy if a norm of reciprocal exchange is operative. Under such circumstances, altruistic behavior can promote individual fitness if the goods received outweigh the costs associated with the helping behavior.

In line with this theory, de Waal (2006; 1996) has documented various forms of prosocial behavior in our closest primate relatives, some of which appear to be rudimentary forms of empathic helping and reciprocal altruism. These behaviors, which de Waal identifies as "proto-moral," will be discussed in greater detail in Chapters 2 and 5. For present purposes, it will suffice to note one of the central implications of this

primate research, as emphasized by de Waal. He underscores that much of the social behavior in primates, including the prosocial actions described above, is thought to be emotionally-mediated (2006, pg. 25). Based on his principle of "evolutionary parsimony"—which posits "if closely related species act the same, the underlying mental processes are probably the same, too" (2006, pg. 62)—de Waal infers that human moral cognition and behavior is also likely rooted in evolved, affective predispositions. Endorsing what he calls an "intuitionist approach to morality," he writes, "I feel that we are standing at the threshold of a much larger shift in theorizing that will end up positioning morality firmly within the emotional core of human nature" (2006, pg. 57).

The empirical research outlined in this section indicates that intuitive ethical judgment is often driven by gut feeling, a finding that undermines Hauser's (2006) cold-processing, Moral Grammar model. Damasio's (2000) VM patient research suggests that somatic markers are necessary for a broad range of moral and social judgments. Haidt and Wheatley's (2005) findings, along with the moral/conventional studies conducted by Nichols (2004a), demonstrate that intense emotional responses can shape our intuitive judgments regarding the severity of moral violations, or even whether a situation qualifies as 'moral' or not. Furthermore, Greene's (2008) fMRI studies show that 'emotional' brain regions are more active in response to moral scenarios involving 'personal harm,' as compared to 'impersonal' situations. Greene's findings may help to explain why VM patients assess 'personal harm' scenarios in an aberrant way, tending toward utilitarian judgments. Finally, research on primate ethical tendencies also supports a sentimentalist account of intuitive ethical judgment. Based on this convergent evidence from a number of fields, it seems that any adequate model of intuitive moral

judgment must incorporate an affective component--and Hauser's (2006) Moral Grammar model should be rejected on this basis.

IV. The Social Intuitionist Model

Haidt and Bjorklud (2008) provide an account of intuitive ethical judgment—the "Social Intuitionist Model [SIM]"--which accords nicely with the research outlined in the previous section. They summarize the two central tenets of their view as follows,

> (1) Moral beliefs and motivations come from a small set of intuitions that evolution has prepared the human mind to develop; these intuitions then enable and constrain the social construction of virtues and values, and (2) moral judgment is a product of quick and automatic intuitions that then give rise to slow, conscious moral reasoning (pg. 181).

With regard to the second central tenet, according to the SIM, intuitive moral assessment occurs when the perception of a moral situation (e.g., 'X is a case of murder') automatically triggers an affective response (e.g., an aversive feeling), leading directly to a corresponding assessment (e.g., 'X is wrong'). In line with the 'moral dumbfounding' findings outlined in Section I, Haidt and Bjorklund characterize ethical 'reasoning' as an "an effortful process usually engaged in after a moral judgment is made, in which a person searches for arguments that will support an already made judgment" (pg. 189). According to this account, moral 'reasoning' is typically just a means of rationalizing our intuitive biases.

The general picture of moral cognition sketched by Haidt and Bjorklund is nearly identical to Greene's model of deontological-style judgment and reasoning. To recall,

Greene (2008) postulates that deontological judgments are determined by emotional alarm bells, and our 'reasoning' regarding these intuitive assessments generally amounts to little more than *post hoc* rationalization. The major difference between these largely complementary theories is that Greene's "dual processing" model distinguishes between two psychological types of judgment, intuitive and reason-based, based on the distinctive type of emotion characteristically linked to each class of assessment. Accordingly, Greene acknowledges that conscious moral reasoning can sometimes play a more determinate role in moral judgment—i.e., in the case of reason-based assessments, which are more subtly influenced by somatic markers, as opposed to emotional alarm bells. Haidt and Bjorklund, on the other hand, do not distinguish between differing psychological types of judgment, suggesting instead that the SIM applies across the board. Haidt and Bjorklund emphasize that "moral judgment should be studied as a *social* process, and in a social context moral reasoning matters" (pg. 193); but, on their view, social 'reasoning' is just a form of rhetorical persuasion, aimed at triggering emotionally-laden intuitions. They write, "the reasons that people give to each other are best seen as attempts to trigger the right intuitions in others…Rhetoric is the art of pushing the ever-evaluating mind over to the side the speaker wants it to be, and affective flashes do most of the pushing" (pg. 192).

The SIM works well as a basic model of intuitive moral judgment and the reasoning-style (i.e., *post hoc* rationalization) to which it is commonly linked, in accord with the empirical evidence cited above regarding moral dumbfounding and the role of gut feelings in assessments of this type. In my opinion, however, Haidt and Bjorklund over generalize the model. As noted above, in chapter 3, I endorse the basic

psychological distinction Greene draws between intuitive judgments and reason-based assessments. The SIM does not account for the more nuanced, 'open' type of reasoning characteristic of the latter class of judgment. Nor does the SIM capture the complex types of moral *normative* reasoning in which philosophers, for example, engage (see Chapter 5). Despite these limitations, the SIM, nonetheless, is an empirically plausible account of *intuitive moral judgment*, which is defined in this project as a psychological class of ethical assessment that is relatively automatic, limitedly based on conscious moral reasoning, and characteristically linked to intense moral feeling.

Recall that, according to the first tenet of the SIM, our moral intuitions have an evolutionary origin. Haidt and Bjorklund (2008) theorize that we come equipped with a set of evolved affective predispositions, from which many of our ethical intuitions emerge. They identify five basic clusters, which they also loosely refer to as "modules," of emotional biases relevant to moral cognition: harm/care, fairness/reciprocity, in group/loyalty, authority/respect, and purity/sanctity. As characterized by Haidt and Bjorklund, the harm/care module incorporates an innate aversion to suffering, as well as an inborn tendency to respond positively to signs of affection. The fairness/reciprocity domain encompasses "a set of emotional responses related to playing tit-for-tat, such as negative responses to those who fail to repay favors" (pg. 203). Included as part of the in group/loyalty module is an inherent tendency to feel favorably disposed to groups that an individual identifies with and to become angry with traitors; while the authority/respect dimension reflects "a set of concerns about navigating status hierarchies, e.g., anger toward those who fail to show proper signs of deference and respect" (pg. 203). Finally, the purity/sanctity domain includes feelings of disgust towards objects that are identified

as impure or contaminated. Haidt and Bjorklund write, "these five sets of intuitions should be seen as the foundation of intuitive ethics. For each one, a clear evolutionary story can be told and has been told many times" (pg. 203). As discussed below, I remain agnostic regarding the accuracy of Haidt and Bjorklund's classification system (e.g., that our evolved intuitions fall into five basic categories), while endorsing their general claim that many of our intuitive ethical judgments are driven by evolved, affective bias.

According to Haidt and Bjorklund, the five general clusters of affective predisposition they identify serve as basic building blocks and constraints for our socially-constructed moral systems. They write, "each of our five foundations can be thought of either as a module itself, or, more likely, as a 'learning module'—a module that generates a multiplicity of specific modules during development within a cultural context" (pg. 205). In loosely identifying these clusters of evolved affective biases as "modules," Haidt and Bjorklund clarify that they are not referring to a Fodorian-style perceptual module (i.e., an informationally encapsulated, domain-specific processing system):

> modules for higher cognition do not need to be as tightly modularized as Fodor's perceptual models…There can be many bits of mental processing that are to some degree module-like. For example, quick, strong, and automatic rejection of anything that seems like incest suggests the output of an anti-incest module, or modular intuition (pg. 205).

They further indicate that these affectively-driven modules furnish the framework and constraints, while culture determines the particular content of moral codes. They underscore, "no culture can construct virtues that do not mesh with one or more of the

foundations….[however], the five foundations greatly underspecify the particular form of the virtues and the constellation of virtues that will be most valued" (pg. 209). According to this view, moral systems or rules that run counter to evolved intuitions will not stand the test of time. Our affective biases, however, leave substantial room for cultural variability. For instance, Haidt and Bjorklund note that some cultures moralize objects of disgust, making the purity/sanctity dimension a central part of their moral system (e.g., the need to eat Kosher in Orthodox Judaism), while other societies view this as a matter of mere convention. More generally, for any given affective module, culture helps to determine the situational triggers. For example, we might be hardwired to respond favorably to perceived cases of 'kindness,' but what constitutes an act of kindness may vary from culture to culture, and this local teaching will tune our harm/care module accordingly.

While the SIM presented by Haidt and Bjorklund (2008) is *similar* to the Moral Grammar view endorsed by Hauser (2006) *in so far as both accounts allow for substantial cultural determination of biological endowment*, there are important differences between these views. Chandra Sripada (2008) provides a helpful framework for distinguishing these theories. Sripada contrasts two general types of nativist ethical accounts, capacity nativism and content nativism. The former concerns cognitive capacities, such as Theory of Mind (see Chapter 2), that are important for ethical cognition, but apply in other domains as well (e.g., everyday social interactions). By contrast, content nativism "concerns the question of whether there is innate structure that shapes the *content of moral norms*, and if there is, what is the nature of this innate structure" (pg. 322). As defined by Sripada, "the content of a moral norm consists of the

class of actions that the norm prohibits, permits or requires" (pg. 321). Offered to explain the same general phenomenon—"the manner in which moral norms exhibit both commonalities and differences in content across human groups" (Sripada, 2008, pg. 322--the SIM and Moral Grammar model of intuitive moral judgment exemplify distinctive versions of content nativism.

Sripada outlines three types of content nativist views. A *Simple Innateness* model "proposes that humans possess an innate body of moral rules and principles…[arising] without the need for any highly specific instruction or cultural inputs…" (pg. 320). As Sripada emphasizes, this type of content nativist theory has trouble accounting for the diversity of moral norms exhibited across cultures. By contrast, a *Principles-Parameters* model, like Hauser's (2006) Moral Grammar theory, allows for the cultural specification of innate principles, principles which circumscribe the range of possible norms. According to Sripada, this view accounts for cultural variability in moral norms based on "the operation of universal, underlying moral principles that allow for a highly restricted range of parameterized variability" (pg. 326). Hauser et al. (2008) explicitly refer to their model in terms of a 'principles-parameters' framework. For example, they write, "the hypothesis here is simple: our moral faculty is equipped with a universal set of principles, with each culture setting up particular exceptions by means of tweaking the relevant parameters" (pg. 122).

Sripada argues that a *Principles-Parameters* model is inadequate. He writes, "the pattern of variation in moral norms cannot be explained in terms of the operation of a few relatively rigid parameters" (pg. 329). Instead, Sripada endorses an *Innate Biases* model, the sort of content nativist theory Haidt and Bjorklund (2008) propose. According to this

more modest account, we do not possess innate moral *principles*. Rather, we have evolved *affective predispositions* that incline us towards endorsing some moral norms, while rejecting others. Sripada writes,

> An 'innate bias' on the contents of moral norms is some element of innate structure that serves to make the presence of some moral norms…*more likely* relative to the case in which the bias is absent….However, unlike in the case of the Principles and Parameters model, which involved more or less impermeable parameters, an innate bias does not *require* or *preclude* the presence of any particular moral norm or set of moral norms (pg. 332).

Sripada contends that, of the three types of content nativist views, the *Innate Biases* variety has the most empirical support. Echoing Haidt and Bjorklund, Sripada writes, "I believe the best description of the pattern of variation in moral norms is…'thematic clustering.' There are certain high-level themes that one sees in the contents of moral norms in virtually all human groups" (pg. 330). Sripada emphasizes that, nonetheless, "the *specific rules* that fall under these high-level themes exhibit enormous variability" (pg. 330). Hence, according to Sripada, an *Innate Biases* model, as opposed to a *Principles-Parameters* view, allows for more flexibility in the content of moral norms. We do not possess innate general *rules*, with culture determining the exceptions. Rather, we exhibit affective biases, subject to cultural tuning. While lacking the determinate structure of a universal rule, these affective predispositions lead to pan-cultural normative concerns.

Sripada correctly endorses the *Innate Biases* model, but he does so for the wrong reason. In making his case, Sripada repeatedly emphasizes the rigidity of the *Principles-*

Parameters position, stressing that parametric variability, with the limited range of possibility it entails, is overly restrictive. It is not clear, however, that parametric variability necessarily involves the operation of only "a few relatively rigid parameters." Indeed, in presenting their view, Hauser et al. (2008) do not emphasize parametric rigidity. On the contrary, they indicate that parameters (i.e., the range of exceptions to universal rules) are highly variable, dependent on local teaching. The debate between these two content nativist views comes down to whether we possess innate general *rules*, as opposed to weaker affective *biases* organized into thematic clusters. I believe both theories are equally capable of accounting for cross-cultural similarities and differences in the content of moral norms. The central reason to prefer an *Innate Biases* model is that it accords better with the evidence outlined in Section III regarding the role of emotion in intuitive ethical judgment. Hauser's Moral Grammar model leaves emotion out of the picture, suggesting instead that our snap assessments are based on the cold-processing of parameterized principles, innate rules having no clear evolutionary precursor. In this regard, Haidt and Bjorklund's (2008) *Innate Biases* account is clearly superior.

In outlining their SIM, Haidt and Bjorklund (2008) also address individual moral development. They explain this process in terms of the maturation of evolved affective predispositions tuned by social learning, writing, "moral development can now be understood as a process in which the externalization of five (or more) innate modules meets up with a particular set of socially constructed virtues" (208). They note that the five primary learning modules seem to manifest at differing of stages of development. For example, whereas very young children show sensitivity to suffering in humans and animals (harm/care), a sense of fairness (fairness/reciprocity) and disgust (purity/sanctity)

seems to develop later. Haidt and Bjorklund suggest that, once these basic emotional tendencies come online, local teaching orients them toward specific situations. Children learn from teachers, parents and peers to identify situations as examples of different prototypical moral scenarios. For instance, children are taught what constitutes fair exchange, including exemplary cases and common violations. When they identify a situation as a case of 'fair exchange,' their fairness/reciprocity module 'lights up', leading to a feeling of approbation—but the initial perceptual categorization is based, at least in part, on learning. Outlining this developmental process, Haidt and Bjorklund write:

> The basic idea is that morality, like sexuality or language, is better described as emerging from the child (externalized) on a particular development schedule rather than being placed in the child from outside (internalized) on society's schedule. However as with linguistic and sexual development, morality requires guidance and examples from the local culture to externalize and configure itself properly (pg. 206).

According to this framework, a person's moral awareness and behavior is based in part on individual temperament and distinctive learning experiences (e.g., having better or worse moral instructors and models). Haitdt and Bjorklund speculate that "some people are simply born with brains that are prone to experience stronger intuitions from individual moral modules" (pg. 210). As a result, not all children are equally 'tunable' in each of the five dimensions.

A weakness of Haidt and Bjorklund's (2008*) Innate Biases* theory of intuitive ethical judgment is that the SIM does not include a detailed account of the underlying neuropsychological mechanisms. Paul Churchland's (1996) connectionist account of

prototype representation and processing maps on nicely, however. In *The Engine of Reason, the Seat of the Soul*, Churchland (1996) investigates the cognitive implications of connectionism, providing a neurologically plausible account of conceptual prototypes and the prototype representation of moral knowledge. Connectionism provides an alternative to traditional, symbolic-processing accounts of cognition, which emphasize the rule-based manipulation of symbols, processed in a serial fashion (Bechtel & Abrahamsen, 2002). Connectionist or Parallel Distributed Processing (hereafter, PDP) models, in contrast, are based on the sub-symbolic, simultaneous processing by connected units. In standard versions, each unit has a mathematically expressible "activation level," and is linked to other units via connections with adjustable "connection stregnths," determining the degree of activation transference. Patterns of activation, charted as mathematical vectors, can be represented in "activation spaces." Biologically inspired, these connectionist models provide a very rough approximation of neural processing, and they can be used to simulate various cognitive processes and operations. Within this paradigm, mental representations correspond to patterns of neural firing. Churchland writes, "the general and lasting features of the external world are represented in the brain by relatively lasting configurations of synaptic *connections* (pg. 6)." He then broadly characterizes PDP processing as the "transforming [of] one pattern into another by passing it through a large configuration of synaptic connections" (pg. 11). Finally, he underscores that, in addition to being more biologically plausible, two other notable properties of PDP models of cognition include their relative speed and complex pattern recognition ability (pg. 15).

Churchland bases his account of moral perception on a connectionist theory of conceptual prototypes. According to his model, an activation space simulates a categorical domain, consisting of prototypical representations and other representations variously related to them, each having a distinctive vector. A prototype corresponds to the average vector of all the members in a given category. The closer (mathematically) a representation is to a prototype, the more resemblance they share. To visualize the idea, Churchland asks us to envision a three-dimensional, activation cube. At the very center is a prototype representation, with other members variously situated in relation to it. The greater the proximity to the center, the more the representation resembles (i.e., shares features with) the prototype. Category groupings and formations can be adjusted through altering connection weights, simulating the refinement of synaptic connections through learning. Hence, according to Churchland's model of conceptual prototypes, the associated groupings constitutive of these categories are explained in terms of similar patterns of activation based on synaptic connections. The activation of one representation incites the other members of a category, to varying degrees, depending on the level of resemblance (i.e., the degree of positive activation transference). Categories are most readily identified with their prototypes, since these representations generally receive the greatest activation of all the members in the group.

Prototype activation is also a centerpiece of Churchland's explanation of perception, fundamentally conceived as a pattern-recognition task. On this account, sensory inputs are channelled as specific vectors within an activation space, via PDP. The input is then identified or categorized in relation to the prototype triggered by the activation vector. For instance, consider the example of perceiving a cat.

Within the activation space for ‘animal,’ there is a prototype vector of ‘cat.’ In perceiving an animal as a cat, the sensory input triggers an activation vector close to the prototypical representation; which then incites the prototype, allowing for the animal to be identified.

Churchland argues that intuitive moral perception and judgment work in fundamentally the same way—through prototype representation and activation. He criticizes rationalist accounts of morality for overestimating the importance of rule-following; suggesting instead, “it may be that [our] capacity for moral perception, cognition, deliberation, and action has rather less to do with rules, whether internal or external, than is commonly supposed” (pg. 144). Churchland argues, on the contrary, moral cognition depends crucially on the discriminative processing of prototypes. He writes, “the alternative [to rule-based accounts] is a hierarchy of learned prototypes, for both moral perception and behavior, embodied in the well-tuned configuration of a neural network’s synaptic weights” (pg. 144). On Churchland’s account, moral concepts and categories are generated through social experience, in a manner similar to what Haidt and Bjorklund describe. For example, take the prototype ‘greed.’ During childhood, we primarily learn about this concept through engaging in social interactions that require sharing. These early experiences shape our concept of ‘greed,’ which encompasses prototypical examples (e.g. Johnny’s refusal to share his cake) and other less central examples (e.g. dad’s boss didn’t give him a raise). Over time, through further social experience, this category can be enriched and deepened to encompass a wider array of cases, including those that are borderline (e.g. Bill Gates’ unwillingness to give more money to charity), and perhaps different prototypical examples. These representations of

individual cases, which jointly constitute the prototype concept ‘greed,’ are coded and stored as vectors in a “social” activation space. According to this model, intuitive moral judgment stems from the matching of a novel situation to a moral prototype (e.g., ‘X is a case of greed’), which automatically triggers a corresponding judgment (e.g., ‘X is wrong’).

Churchland‘s model does not explicitly incorporate emotion, but his general account of prototype representation and processing can readily accommodate a sentimentalist view of intuitive ethical judgment. What is the connection between moral prototypes and the evaluative judgments they automatically trigger? Churchland indicates that these normative assessments are learned (e.g., in developing a moral prototype for ‘theft,’ a child also learns that people judge such actions to be immoral)—which is consistent with the characteristic, anti-nativist leaning of connectionist theories. From an *Innate Biases* perspective, however, we come equipped with a set of basic moral modules that are emotionally-valenced. These affective predispositions—for instance, finding violence to be generally averse, or feeling sympathy for kin--were adaptive for our ancestors. When first manifested, usually during early childhood, these emotionally-laden prototypes are very general, lacking specific content (i.e., a breadth of real-life examples). Over time, through social experience and learning, of the sort envisioned by Churchland, these prototypes can be developed and enriched, allowing for more nuanced perception. Nonetheless, while moral prototypes can be refined through learning, the affective predisposition connected with each basic type remains consistent. For example, any situation ‘lighting up’ your theft prototype, will trigger a similar aversive response, leading to a negative moral judgment.

There is a tension in the composite account being offered here. While embracing Churchland's connectionist model of prototype representation and processing, I am also endorsing a nativist account of evolved, affective predisposition. Although connectionism is commonly opposed to nativism, attempts have been made to reconcile these two positions (see, for example, Elman et al., 1998), which I believe is a fruitful path to follow. Indeed, *prima facie*, Churchland's account of moral prototypes seems compatible with the *Innate Biases* theory proposed here—and there is good reason to combine the two, since this composite view has the most empirical support. This combined model has the virtue of making Churchland's theory of moral prototypes more consistent with research regarding the evolutionary and affective foundations of ethical judgment, while avoiding the postulation of innate *principles or rules*—which appears to be the primary target of his opposition to nativism. It would be beyond our scope to delve into this debate at great depth. For our purposes here, the central goal was to supplement Haidt and Bjorklund's (2008) SIM of intuitive moral judgment with an empirically plausible account of the underlying neuropsychological mechanisms, in order to further bolster this *Innate Biases* position. In Chapter 5, Paul Thagard's (2006) HOTCO models, which simulate various types of emotional cognition within a connectionist framework, will be discussed in connection with Rawls' method of wide reflective equilibrium.

As noted in the first section of this chapter, Hauser et al.'s (2008) "trolley problem" research is the most wide-ranging study of intuitive ethical judgment to date. This research has revealed recurrent assessment patterns across cultures consistent with an evolutionary genealogy of our moral tendencies. Moreover, the 'moral

dumbfounding' results in this set of studies, as well as those conducted by Haidt and colleagues (2001, 2000, 1993), indicate that conscious moral reasoning generally plays only a limited role in intuitive ethical judgment. For this psychological class of assessment, it appears that moral reasoning often amounts to little more than *post hoc* rationalization. Even when subjects are unable to provide adequate justification for their snap judgments, they tend to remain committed to them, nonetheless.

In light of this evidence, researchers have recently offered new explanatory models of intuitive ethical cognition. The two leading theories, the Moral Grammar Model and the SIM, were outlined in this chapter. According to the former position, intuitive ethical judgments stems from the cold-processing of innate moral principles. It was argued above that the Moral Grammar model lacks empirical support, given the substantial evidence that emotion drives many of our intuitive ethical judgments. As revealed by Koenigs et al (2007)., when given the standard trolley problem test, VM patients, a group with affective-processing deficits, offered aberrant judgments regarding cases involving personal harm, (e.g., the Footbridge scenario), tending toward utilitarian responses. Consistent with this finding, in his fMRI studies, Greene found that, in normal subjects, contemplation of 'personal harm' moral scenarios is associated with relatively greater activation in 'emotional' brain regions, whereas 'impersonal' scenarios (e.g., the Standard trolley problem scenario) elicits higher activity in classically 'cognitive' areas. Nichols' (2004a) moral/conventional studies and Haidt and Wheatley's (2005) findings provide further evidence that intense emotion shapes our intuitive assessments. Nichols' reports that violations of disgust norms—a group of norms linked to strong emotion--and moral norms are judged similarly, in so far as both are characteristically viewed to be

more serious, less permissible, and less authority contingent than violations of conventional norms. Based on these findings, Nichols concludes that, like disgust norms, moral norms are emotionally-laden; and this connection to intense affect is what leads us to attribute greater practical clout to these "sentimental norms." In accord with this hypothesis, Haidt and Wheatley found that emotional manipulation can alter the severity of subjects' snap moral judgments, and even cause some subjects to assess that non-moral situations involve an ethical violation. Finally, as underscored by de Waal (2006), research regarding proto-moral behaviors in our primate relatives also supports a sentimentalist view of intuitive ethical judgment.

In light of this research, Haidt and Bjorklund's SIM is clearly superior to the Moral Grammar alternative. According to Haidt and Bjorklund's *Innate Biases* view, we come equipped with a set of evolved affective predispositions. These broad emotional proclivities are tuned by social learning, which teaches us how to 'appropriately' categorize and respond to morally-salient situations. On this hot-processing account of intuitive moral judgment, the perceptual identification of a moral situation automatically triggers a linked emotional response, which leads directly to a corresponding judgment. Churchland's (1996) connectionist theory of prototype representation and processing provides a neurologically plausible account of the underlying mechanisms.

There are problems with Haidt and Bjorklund's SIM, however. As emphasized above, these theorists fail to differentiate between distinctive psychological types of judgment (e.g., intuitive vs. reason-based) and unique varieties of moral reasoning (e.g., post hoc rationalization versus normative moral reasoning)--important discriminations that will be brought into greater relief in the remaining chapters of this project. While the

SIM works well as a general account of intuitive moral judgment and reasoning, it does not apply across the board. In addition, more research needs to be conducted regarding the five basic domains (e.g., harm/care, in group/loyalty, etc.) of intuitive bias identified by Haidt and Bjorklund. The description they provide of these domains is quite skeletal, and should be fleshed out with a more detailed account of the specific evolved proclivities constitutive of each thematic cluster. For example, in Chapter 5, the evolutionary origins of retributivist, 'eye-for-an-eye' moral principles (e.g., 'one bad turn deserves another'), which would fall under Haidt and Bjorklund's 'fairness/reciprocity' domain, are discussed. While I remain agnostic regarding the accuracy of Haidt and Bjorklund's classification system for our biases (e.g., five basic domains), I endorse their general insight that many of our intuitive ethical judgments are driven by evolved affective predispositions.

This chapter began with a brief discussion of Hume's seminal claim that "reason is, and ought only to be the slave of the passions." As I interpret it, this is fundamentally a statement about the vital role of emotion in moral cognition and the implications for a psychologically realistic approach to normative ethical justification (i.e., we need to attribute normative weight to our gut feelings). This opening chapter has focused on empirical evidence regarding the affective foundations of intuitive moral judgment. The subject of the next three chapters will be other emotionally-laden facets of ethical cognition--pertaining, for example, to empathy and moral motivation—and some of the metaethical implications. This sets the stage for the final chapter in this project in which I outline a psychologically realistic normative reasoning procedure based on Rawls'

method of wide reflective equilibrium. The next chapter focuses on the neuropsychological underpinnings of empathy.

Chapter 1 References

Arnhart, L. (1998). *Darwinian Natural Right: the Biological Ethics of Human Nature.* Albany: State University of New York Press.

Bechtel, W., & Abrahamsen, A. (2002). *Connectionism and the Mind.* Second Edition. Malden, Massachusetts: Blackwell Publishing.

Churchland, P. M. (1996). *The Engine of Reason, the Seat of the Soul*. Cambridge, Massachusetts: MIT Press.

Damasio, A. R. (2000). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Random House.

Damasio, A., Adolphs, R., Tranel, D., & Koenigs, M. (2007). *Get full reference.

De Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge, Massachusetts: Harvard UP.

De Waal, F. (2006). *Primates and Philosophers: How Morality Evolved.* Princeton: Princeton UP.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1998). *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, Mass.: MIT Press.

Greene, J. (2008). The secret joke of Kant's soul. In *Moral Psychology, Volume 3,* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Haidt, J., Kollers, S., & Dias, M. (1993). Affect, culture or morality, or is it wrong to eat your dog? *Journal of Personality and Social Pscyhology* 65: 613-28.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral Dumbfounding: When intuition finds no reason. Unpublished Manuscript, University of Virginia.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108,* 814-834.

Haidt, J, & Hersh, M.A. (2001). Sexual morality: The cultures and reasons of liberals and conservatives. *Journal of Applied Social Psychology, 31,* 191-221.

Haidt, J., & Bjorklund F. (2008). Social Intuitionists Answer Six Questions about Moral Psychology. In *Moral Psychology, Volume 2,* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: Harper Collins.

Hauser, M., Young, L., & Cushman, F. (2008). Reviving Rawls's linguistic analogy: operative principles and the causal structure of moral actions. In *Moral Psychology, Volume 2*, ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Hume, D. [1793] (1964). *A Treatise of Human Nature.* Oxford: Clarendon Press.

Kant, I. [1785] (1964). *Groundwork of the Metaphysics of Morals,* Trans. H.J. Patton. New York: Harper Torchbooks.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446,* 908-911.

Nichols, S. (2004a). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford UP.

Pinker, S. [1971](1997). *How the Mind Works.* New York: W.W. Norton & Company.

Rawls, J. (1999). *A Theory of Justice*. Cambridge: Harvard UP.

Schnall, S., Haidt., J, Clore, G.L., & Jordan, A.H. (2007). Irrelevant disgust makes moral judgments more severe, for those who listen to their bodies. Unpublished Manuscript, University of Virginia.

Sripada, C. (2008). Nativism and moral psychology: three models of the innate structure that shapes the content of moral norms. In *Moral Psychology, Volume 1*, ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Thagard, P. (2006). *Hot Thought.* Cambridge Mass.: MIT Press

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 35-57.

Wheatley, T., & J. Haidt. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science 16*, 780-784.

2. The Two Faces of Empathy: Evidence from Psychopathy and Autism

"No quality of human nature is more remarkable, both in itself and in its consequences, than that propensity we have to sympathize with others…In general we may remark, that the minds of men are mirrors to one another…"

—David Hume, *Treatise*

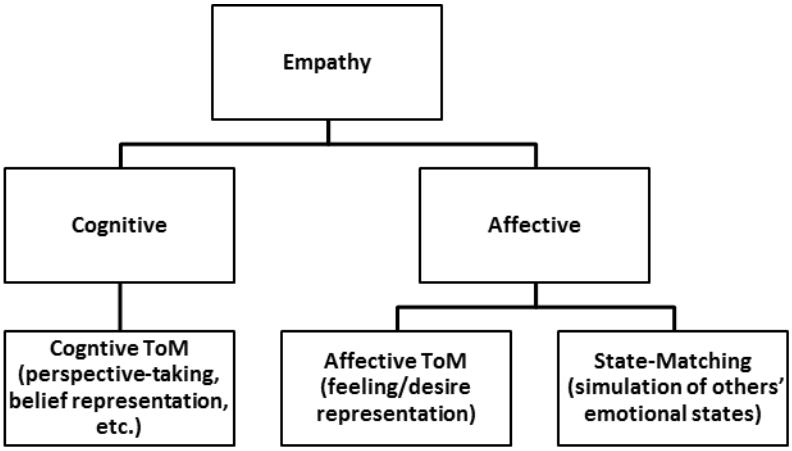
Ethicists in the sentimentalist tradition, like David Hume, have long argued that empathy-related processes are at the core of morality. They emphasize that ethical action in response to the suffering of others requires perception and caring and humans have a remarkable capacity for both. It turns out that these philosophers were right. There are two main facets of empathy, an emotional and a cognitive dimension, as reflected by the distinctive impairments of autists and psychopaths. Psychopaths lack emotional empathy, but their capacity for cognitive empathy is largely intact—a dangerous combination freeing these individuals to pursue callous, premeditated acts of harm. Autists, on the other hand, display profound cognitive empathic deficits and substantial emotional empathic limitations as well, although some rudimentary forms of emotional empathy may be present in this population. As a result, they are prone to display a different type of moral shortcoming--a form of ethical *neglect* involving a failure to respond with sensitivity to the needs of others. Although the respective moral failings of autists and psychopaths differ, each characteristic shortcoming demonstrates the ethical importance of having both dimensions of empathy function together.

When it comes to understanding the underlying neuropsychological mechanisms, however, distinguishing between these two aspects of empathy is helpful. Unfortunately, in much of the empathy literature this has not been standard practice. Summarizing this predicament, Batson (2009) writes, "although [students of empathy] typically agree that empathy is important, they often disagree about why it is important, about what effects it

has, about where it comes from, and even about what it is" (pg. 3). Accordingly, terminological confusion abounds in the literature. Batson has identified no less than eight distinct empathic achievements, ranging from basic motor mimicry to imaginative role-reversal, all of which have been termed 'empathy' by various authors. Addressing a source of the problem, Batson writes, "application of the term empathy to so many distinct phenomena is, in part, a result of researchers invoking empathy to provide an answer to two quite different questions: How can one know what another person is thinking and feeling? What leads one person to respond with sensitivity and care to the suffering of another (pg. 3)?" Each of these questions--the first concerning how we understand the mental states of others and the latter dealing with our sympathetic response to this awareness—roughly corresponds to one of the two main dimensions of empathic processing, cognitive and emotional, respectively. The correspondence is only approximate because there appears to be two (partially) separable systems involved in the understanding of mental states, an affective Theory of Mind (ToM) network and a cognitive ToM network. The former deals primarily with the representation of basic feelings and desires, whereas the latter involves self-other differentiation mechanisms, conscious perspective-taking, and the representation of belief. As defined here, 'cognitive empathy,' one of the two main facets of empathic processing, refers to cognitive ToM processes. By contrast, I am including affective ToM under the general 'emotional empathy' category because it appears to incorporate similar neuropsychological mirroring systems to those involved in emotional state-matching. Emotional state-matching refers to cases in which the representation of another's

affective state automatically elicits a *similar*, but not necessarily identical, response in the observer.

Figure 1: The Two Facets of Empathy



In Sections I and II, based on autism and, especially, psychopathy research and independent neurological findings, it is argued that cognitive and emotional empathy are at least partially dissociable at a neuropsychological level; a finding which may help to resolve a hot philosophical debate concerning our ToM capacities. "Simulation theorists" argue that we understand the mental states of others by internally recreating these observed states, while "theory theorists" contend that we generate hypotheses, without actually simulating the states of others. It appears that both camps are partially right. Affective ToM incorporates relatively more simulation-based mechanisms, while cognitive ToM involves more 'theoretical' operations. Evolutionary evidence also supports the basic distinction between these two dimensions of empathy. As discussed in Section III, Frans de Waal (2009) proposes a Darwinian, "Russian Doll" model, with emotional empathic processes at the core. Finally, in the last section of this chapter, I briefly outline how a similar approach to the one adopted here for the analysis of empathy-- a method distinguishing between its more emotional and cognitive

dimensions--may be usefully applied to the study of moral conscience and the related emotions of shame and guilt.

I. Emotional Empathy and Psychopathy

Of all the personality disorders, psychopathy is one of the most frightening. Psychopaths display a blatant disregard for the welfare of others and a chilling lack of guilt or remorse for their misdeeds. Addressing these deficits, Robert Hare (1993) writes,

> Psychopaths show a stunning lack of concern for the devastating effects their actions have on others. Often they are completely forthright about the matter, calmly stating that they have no sense of guilt, are not sorry for the pain and destruction they have caused, and that there is no reason for them to be concerned…[they] display a *general* lack of empathy. They are indifferent to the rights and suffering of family members and strangers alike (pg. 41, emphasis in the original).

Hare characterizes these impairments as a "general lack of empathy." A more precise description, however, is that psychopaths have emotional empathic deficits, specifically with regard to sympathy-related processing. Indeed, James Blair and Karina Blair (2009) argue that this type of emotional dysfunction accounts for this population's lack of moral socialization. By contrast, as evinced by their characteristic manipulative actions and deceitfulness—behaviors which require impressive perspective-taking abilities--psychopaths do not appear to have substantial cognitive empathic limitations.

As noted above, emotional state-matching occurs when the perception of another's emotional state generates a *similar* feeling in the observer; for example, feeling

happy in the company of a laughing friend. Cases in which there is a very close correspondence between perceived and mirrored affect are typically referred to in the literature as examples of "emotional contagion." The prevailing view among neuroscientific empathy researchers is that emotional state-matching rests on a simulation-based processing system, whereby the perception of another's affective state *automatically* triggers a similar state in the observer. Shimon Shamay-Tsoory (2009), for example, writes, "affective empathic response is driven mainly by simulation, involving regions that mediate emotional experiences (i.e., the amygdala, insula and the inferior frontal gyrus)" (pg. 228). There is greater controversy concerning the question of whether there are separable networks in this mirroring system for the processing of distinctive emotions, such as fear and anger (Blair and Blair, pg. 140). As I will outline below, psychopathy research supports this thesis. Our primary focus in this section, however, will be state-matching responses to the detection of pain and suffering in others—an emotional empathic process--which is especially relevant to morality. Two common responses, personal distress and sympathy, are both lacking in psychopaths, which may help to explain their immoral behavior.

Before turning to these two specific types of responses, it will be helpful to consider in greater detail the general 'mirroring' model of emotional state-matching, which now predominates. Several theorists, including Decety and Jackson (2004) and Preston and de Waal (2002), have endorsed variants of this "perception-action" view. Preston and de Waal, for instance, propose a "Perception Action Model (PAM)" of empathy, which stipulates, "[the] attended perception of the object's state automatically activates the subject's representation of the state, situation and object, and that activation

of these representations automatically primes or generates the associated autonomic and somatic responses, unless inhibited" (4). According to this account, perceiving the emotion of another can automatically incite a similar feeling in the observer. The reason for this correspondence is that the emotions of self and other are neurologically represented in the same way. For example, my sadness and your sadness activate the same 'sadness' network in my brain, a distributed representation which includes somatic and motor connections. Decety and Lamm (2009) characterize this theory of "shared representation" as follows, "perceiving someone else's emotion and having an emotional response, or subjective feeling state, both draw upon essentially the same computational processes and rely on somatosensory and motor representations (pg. 200).

A major breakthrough in support of simulation-based theories of emotional empathy came from the recent discovery of mirror neurons. These special classes of neurons are coded for particular actions, firing either when the action is directly executed or merely observed--for example, grasping an object or watching someone else perform a similar action. Mirror neurons were first detected by a research group in Parma using singe-cell recordings in monkeys (Thagard, pg. 188). Similar types of neurons connected with the processing of pain and disgust have already been identified in humans (Thagard, pg. 188). For example, experiencing disgust and observing disgusted facial expressions in others activates similar regions of the insula (Decety and Lamm, pg. 201). Emphasizing the significance of these findings in relation to perception-action models of emotional empathy, Decety and Lamm write, "the discovery of sensorimotor neurons (called mirror neurons) in the premotor and posterior parietal cortex that discharge during both the production and the perception of the same action performed by another

individual provides the psychological mechanism for [a] direct link between perception and action" (pg. 200). Drawing a similar conclusion, Preston and de Waal emphasize, "[mirror neurons] do provide concrete cellular evidence for the shared representation of perception and action" (pg. 10). Again, the guiding idea of perception-action models of emotional empathy is that, unless inhibited, the perception of another's emotion automatically incites a similar emotional response in the observer, based on a shared neural representation for the emotions of self and other. Mirror neurons appear to provide a mechanism for this shared representation.

As reflected by psychopathy, lacking an emotional responsiveness to the pain and suffering of others in conjunction with an intact capacity for cognitive empathy, may lead to very bad moral outcomes. Two related types of state-matching response, sympathy and personal distress, are especially relevant in this regard (Decety and Lamm). Both involve an aversive reaction to the perceived distress of others and can motivate helping behaviors. There is, for instance, a vast range of psychological research linking sympathy to charitable assistance and positive relationship outcomes (Tagney and Darley, pg. 79). Personal distress, in comparison, is a relatively more self-focused emotion, which can lead to withdrawal, as opposed to solicitous action. While perhaps a less morally desirable response to the suffering of others than sympathy, personal distress may play an important role in violence-inhibition. As discussed below, Blair and Blair argue that psychopaths' lack of emotional responsiveness to the suffering of others—i.e., their failure to experience personal distress and sympathy—frees them to act in violent ways. With regard to the underlying processes, Decety and Lamm report that, consistent with perception-action models of emotional state-matching, personal distress and

sympathy stem from mirroring-based neural activity. They write, "current neuroscientific evidence suggests that merely observing another individual in a painful situation yields responses in the neural network associated with the coding of the motivational-affective dimension of pain in oneself" (pg. 201). Decety and Lamm argue that the perceived suffering of another automatically elicits personal distress in the observer, and this self-focused response must be regulated in order for sympathy to emerge. As an 'outward looking' emotion, sympathy requires attention to another's distress, and not just one's own. Accordingly, too much personal distress can block sympathetic responsiveness, which requires a greater degree of perspective-taking and self-other differentiation. Decety and Lamm write,

> if perceiving another person in an emotionally or physically painful circumstance elicits [a high degree of] personal distress, the observer may tend not to fully attend to the other's experience, and as a result may fail to display sympathetic behaviors…Taking the perspective of the other produces additional activation in specific parts of the frontal cortex that are implicated in executive functions, particularly inhibitory control…This ability is of particular importance when observing another's distress, because a complete merging with the target would lead to a confusion as to who is experiencing the negative emotions…"(pg. 204).

Paul Thagard (2010) nicely encapsulates the moral relevance of pain-related mirroring systems. A problem for moral psychology is figuring out why people care about the pain and suffering of others. It makes sense that people should find their own suffering aversive, but what psychological mechanisms underlie our sympathy for others and why is this natural concern missing in psychopaths? The answer, it seems, is that

normal individuals experience the pain of others as if it were their own, based on simulation processing. Addressing this connection, Thagard writes,

> Mirror neurons provide the plausible missing link between personal experience and the experience of others. People not only observe the pain and disgust of others; they experience their own versions of that pain and disgust, as shown by the mirroring activity in cortical regions such as the insula and anterior cingulate cortex…Normal children do not need to reason about why harm is bad for other people; they can actually feel that harm is bad. Thus mirror neurons provide the motivation not to harm others by virtue of direct understanding of what it is for another to be harmed (pg. 194).

In accord with this theory, a plausible explanation for the lack of sympathy in psychopaths is that they have dysfunctional mirroring networks for the processing of pain and suffering (194). As outlined below, this group shows clear deficits in responding to the fear and distress-cues of others, which appears to stem from amygdala abnormalities. Accordingly, Thagard speculates, "it is possible that psychopaths' deficits in emotional learning that involve disrupted functioning of the amygdala are partly due to mirror neuron malfunctioning" (pg. 194).

Consistent with the view that personal distress and sympathy in response to the perceived suffering of others relies on similar neuropsychological pathways, both reactions appear to be lacking in psychopaths. James Blair, Derek Mitchell and Karina Blair (2003) report a variety findings linking psychopathy to dysfunctional processing of fear and distress (pg. 54). For instance, as compared to normals, psychopaths show reduced autonomic responses (as measured by skin conductance) to the perceived distress

of others. Psychopaths also display impairments for the recognition of fearful and sad expressions, while performing at normal levels for angry, happy, and surprised expressions. This selective impairment supports a view that separable emotional mirroring networks handle unique types of expressions, a thesis that Blair & Blair endorse (2009, pg. 140). For instance, noting that the processing of fear, anger and disgust have been linked to distinctive brain regions (amygdala, orbital frontal cortex, and insula, respectively), they write, "it is unlikely that a unitary system responds to all expressions" (pg. 140). With regard to the issue at hand, fear and distress-processing deficits in psychopaths, members of this population also show reduced levels of fearfulness and a decreased level of responsiveness to threatening stimuli. Summarizing the findings from various studies, Blair et al. (2003) write, "[psychopaths] show reduced aversive conditioning, reduced emotional responses in anticipation of punishment, reduced emotional responses when imagining threatening events, and reduced augmentation of the startle reflex by aversive primes (pg. 50)."

This finding that the abnormal processing of fear and distress in psychopathy impacts both self-experience and other-directed perception is consistent with mirroring-based explanations of emotional state-matching. Again, according to the general perception-action theory, perceiving another's emotional state activates similar neural pathways to the ones involved in the first-person experience of this emotion. Accordingly, one plausible explanation for a failure to perceive in others a particular type of emotional expression, such as fear, is damage to the underlying network responsible for this type of affective-processing; which, in turn, could limit the first-person experience of this emotion as well. Indeed, Blair & Blair argue that the general distress-

processing deficits in psychopathy, which disrupts both the first-person experience of these emotions as well as empathic responses (i.e., personal distress and sympathy) to the perceived suffering of others, stems from amygdala abnormalities, an area of the brain linked to fear-based processing. They emphasize, "considerable data suggest amygdala dysfunction in psychopathy" (pg. 144). According to this view, there are separate networks in the mirroring system responsible for different types of emotional expressions (fear, anger, happiness, etc.). The network responsible for the simulation (and first-person experience) of fear and distress is abnormal in psychopaths, which explains their selective emotional empathic deficits relating to these particular emotions.

Blair & Blair (2009) contend that these affective impairments can, in turn, explain the lack of moral socialization characteristic of psychopathy. They write,

> Individuals with psychopathy show clear impairment in processing the fearfulness and sadness of others…According to the argument developed here, the important empathic process with respect to moral socialization is the translation of the victim's distress such that stimulus reinforcement occurs; the victim's distress is aversive to the healthy individual, who learns to avoid the action that has caused the other harm (pg. 144).

Blair & Blair propose that normal individuals naturally find distress cues aversive, and as such learn to avoid behaviors (e.g., hitting others) that elicit these responses. Psychopaths, in contrast, who are devoid of this type of emotional empathy, cannot be conditioned in this way; which, in turn, permits them to pursue instrumental violence. As opposed to reactive aggression, which is a more spontaneous response to frustrating or threatening events, instrumental aggression is more purposeful and goal-directed (e.g.,

robbing someone for monetary gain, etc.). Normal individuals learn to avoid these behavioral strategies because they cause others to suffer (an aversive stimulus). Without these natural controls, psychopaths are unable to deeply internalize the moral rules that proscribe such behaviors—and their capacity for cognitive empathy allows them to successfully plan and execute these acts of instrumental aggression. Blair & Blair underscore that, consistent with this account, psychopaths fail to distinguish between moral and conventional normative violations; which Blair & Blair contend stems from this population's insensitivity to the distress cues that are characteristically linked to moral norms.

Blair and Blair argue that autists, on the other hand, do not typically display this specific type of emotional empathic impairment. They report that, unlike psychopaths, individuals with autism--at least those who are higher functioning--possess a rudimentary sensitivity to the suffering and distress of others and they typically pass the moral-conventional distinction test. Blair & Blair write,

> [there is evidence] that children with autism show autonomic responses to the distress of others and that at least those who are more cognitively able are appropriately emotionally responsive to the distress of others. In short, there are reasons to believe that the basic emotional empathic response—that is, the engagement of emotional learning systems following the presentation of emotional expression—is intact in individuals with autism (pg. 147).

Blair & Blair also emphasize that, in comparison to psychopathy, the empathic deficits characteristic of autism are more cognitive in nature, involving a failure to understand the beliefs and intentions of others (pg. 146). As discussed in the next section, there is some

controversy regarding the emotional empathic capacities of autists. In *Mindblindness*, Simon Baron-Cohen (1997) does not directly address this question; but he indicates that autists may possess some very basic emotional empathic abilities (e.g., a capacity to recognize simple emotions, like happy, sad, etc.), while emphasizing their profound cognitive empathic limitations. In his most recent book, however, he (2011) describes individuals suffering from "classic" autism as lacking both emotional and cognitive empathy, and contrasts them to psychopaths. He writes,

> the psychopath *is* aware that he is hurting someone because the 'cognitive' (recognition) element of empathy is (largely) intact, even if the "affective" element (the emotional response to someone's else's feeling) is not. The person with classic (low functioning) autism often lacks both of these components of empathy (120).

This controversy will be addressed in greater detail in the next section. In this section, I delineated the general perception-action theory of emotional state-matching, according to which the observation of another's emotional state automatically generates a similar state in the observer, unless inhibited, based on shared networks of representation for the emotions of self and other. This general model provides a compelling neuropsychological account of the more emotional sides of empathy—such as the personal distress and sympathetic concern that can automatically follow from the perception of another's suffering. These natural sympathies, in turn, motivate a variety of ethically-desirable behaviors, e.g., donating money to charity or helping a person in distress. In addition, as evidenced by the immorality of psychopaths, this type of affective responsiveness also provides an important safeguard against violent and callous

behavior. It appears that normal individuals avoid hurting others because, in an important sense, they share their pain; based on a mirroring network for fear and distress-processing in the amygdala, which is dysfunctional in psychopathy and perhaps in autism as well. Without this natural aversion, psychopaths are psychologically unburdened by the suffering of others.

There is more to empathy, however, than simply feeling. Perception-action models provide a good account of how emotional reactions typically *follow* from the internal representation of another's mental state, as well as how feelings can be automatically represented (an affective ToM process) through direct observation. But in many cases there are more cognitive facets to empathic representation and action, which appear to require specialized neuropsychological mechanisms in addition to perception-action processes. Making a similar point, Batson argues, "to claim that either neural response matching or motor mimicry is the unifying source of all empathic feelings seems to be an overestimation of their role, especially among humans" (pg. 5). The key issue here is *how* we come to represent the mental states of other people, representations which may then serve as inputs to the perception-action networks for emotional empathy. Although the mirroring system described above appears to play an important role in affective ToM processing (e.g., the representation of simple feelings and desires), other neuropsychological mechanisms appear to underlie the more cognitive aspects of ToM (e.g., understanding belief and intention, consciously adopting another's perspective, etc.). In the next section, the general distinction between emotional and cognitive empathic processing will be further illuminated in light of autism research. As noted above, there are questions regarding the degree of emotional empathic limitation in

autists. By contrast, there is no doubt that cognitive ToM capacities are severely limited in this population, whereas cognitive empathy is typically intact in psychopathy.

II. Cognitive Empathy and Autism

Autism is a developmental disorder characterized by severe social impairments emerging early in development. These deficits typically include a lack of social awareness, inappropriate social behavior, a failure to make eye contact, one-sided interaction styles and difficulties with group assimilation (Baron-Cohen, 1997, pg. 62-63). The severity of these symptoms can vary widely and may be exacerbated by a variety of other disorders (e.g., mental handicap, epilepsy, etc.) and for this reason autism is typically characterized as a "spectrum disorder" (Baron-Cohen, 1997, pg. 60). As a result, it is difficult to generalize about standard impairments in moral cognition and behavior in autism, and there is a limited range of literature focusing specifically on this topic. In general, the ethical failings that do emerge typically involve a special type of moral *neglect*, as opposed to the premeditated acts of violent harm perpetrated by psychopaths. Autists can sometimes appear cold and indifferent, failing to appropriately respond to the needs of others. They have been described as treating people more like objects, ignoring their inner life. In other ways, however, autists can appear quite moral, especially as compared to psychopaths. Although individuals with autism may display reactive aggression in response to frustrating events, they are not prone to instrumental aggression. Indeed, as discussed below, individuals with Aspberger Syndrome, on the higher functioning end of Autism Spectrum Disorder, often profess deep ethical

concerns, approaching morality in a strict, Kantian fashion (i.e., a rigid, rule-based orientation).

What accounts for the unique moral profile of autists? Why do they display such different behaviors from psychopaths, another group with profound empathic deficits? As noted in the previous section, Blair and Blair (2009) propose that autists, at least those who are higher functioning, have an intact capacity for emotional empathy while suffering from profound cognitive empathic deficits, which is a reversal of the pattern in psychopathy. Accordingly, Blair and Blair argue that this ability to respond to distress cues in others—an emotional empathic response—allows autists to be morally socialized (i.e., learn to avoid behaviors that cause others pain). As Baron-Cohen (2011) contends in his most recent book, however, the picture is likely more complicated than what Blair and Blair suggest. It is currently unclear to what degree emotional empathic capacities are preserved in individuals with autism, but there are clearly major limitations. For this reason, Baron Cohen (2011) attributes their capacity for moral concern and lack of instrumental aggression to a more general proclivity for "systematizing thought." This will be addressed in greater detail below. What remains beyond doubt is that cognitive empathy is profoundly disrupted in autism, which helps to account for the distinctive type of moral impairment (i.e., a failure to respond with sensitivity to the needs of others) characteristic of this population.

This explanation accords with Baron-Cohen's (1997) diagnosis in his earlier book, *Mindblindness.* In this text, although he does not explicitly distinguish between cognitive and affective ToM networks, Baron-Cohen indicates that autism involves more pronounced deficits in the former, while indicating that basic elements of the affective

ToM network may be preserved. In developing his theory, he argues that the ToM processing system in normal humans involves four main components: an Intentionality Detector (ID), an Eye-Direction Detector (EDD), a Shared-Attention Mechanism (SAM) and a Theory of Mind Mechanism (ToMM). Baron-Cohen links ID to affective ToM processes, while indicating that the other three components of the ToM system are more involved in cognitive ToM operations. According to his model, ID works as an agency-detection device, identifying characteristic types of motion (e.g., self-propelled) as volitional in nature while interpreting these movements in terms of desires and goals (pg. 33). Experimental findings indicate that autists perform close to the level of normal subjects on tests of ID functioning. Baron-Cohen reports, for instance, that autists spontaneously identify actions in terms of desires and goals, distinguish animate from inanimate objects, and can understand simple connections between desires and emotions (e.g., 'getting what you desire leads to happiness'). Autists, nonetheless, characteristically fail to comprehend more complex emotions and intentions, which Baron-Cohen attributes to deficits in their SAM and ToMM networks. He writes, "I suggest that ID is probably functioning normally in children with autism…[which] does not mean that they are able to understand all aspects of desire, or the more complex mental state of intention" (pg. 63).

Baron-Cohen suggests that that the Eye-Direction Detector (EDD) operates normally in autists as well. Describing its main function, he writes, "EDD interprets stimuli in terms of what an agents sees" (pg. 39). According to this model, EDD identifies the presence of eye-like stimuli and determines the object of their gaze. Based on findings that autists readily interpret eye direction in terms of 'seeing' and can

determine what others are looking at, Baron-Cohen argues that this component his four-part ToM system is also intact in this population.

According to his diagnosis, the "mindblindness" exhibited by autists stems primarily from Shared-Attention Mechanism (SAM) and Theory of Mind Mechanism (ToMM) dysfunction. Indeed, Baron-Cohen emphasizes the limited ToM capacities afforded by the Intentionality Detector (ID) and Eye Detection Device (EDD), noting that they jointly allow only for "dyadic" representation (i.e., Agent-Object, Agent-Self), which is insufficient for the representation of shared experience. Emphasizing the limitations of ID and EDD, Baron-Cohen writes, "these mechanisms do not allow you to represent that you and someone else…are both attending to the same object or event. And yet that is exactly what one would need in order to communicate about a shared reality and to feel that you and the other person are focusing on and thinking about the same thing (pg. 44)." He proposes that this type of mutual experience requires joint attention and "triadic" representation (Agent-Self-Object) capacities, which he attributes to SAM. He argues that this vital mechanism is either entirely absent or very late to develop in autists, as shown by their lack of joint-attention behavior (pg. 66). He speculates that SAM interfaces with both ID and EDD to link eye-direction to an agent's goals and desires. For instance, incorporating input from the ID and EDD systems, SAM allows for an instinctual inference that 'agent *wants* the chocolate bar' based on the observation 'agent is looking at the chocolate bar.' Autists who have a dysfunctional SAM network are unable to draw this connection between seeing and desiring. Although they can recognize that an agent is *looking* at a chocolate bar, they do not infer on this basis that the agent *wants* the chocolate bar; even though they are capable of recognizing

simple desires by other means via the ID network. For example, they should be able to infer that the agent wants the chocolate bar if they can observe the agent picking it up.

The final, and perhaps most central component, of Baron-Cohen's four-part ToM system, which incorporates both affective (i.e., the Intentionality-Detector) and cognitive ToM components, is the Theory of Mind Mechanism (ToMM). According to his account, ToMM, a *cognitive* ToM component, allows for the representation of a full range of mental states (thinking, believing, pretending, etc.), building on the basic triadic representations funded by the Shared-Attention Mechanism (SAM). Baron-Cohen speculates that ToMM represents mental states in the form of propositional attitudes (e.g., Mary *thinks* music is wonderful). This propositional form, in turn, allows for "referential opacity," defined by Baron-Cohen as "the property of suspending normal truth relations of propositions" (pg. 52). The basic idea is that 'intentional' propositions, such as "Mary believes we live on the planet Mars," may be true (i.e., Mary really believes this), even though the component proposition ('we live on the planet mars') is false. Baron-Cohen suggests that normal children by around four to five years of age acquire a firm, intuitive grasp of referential opacity—which is demonstrated by their capacity to understand that people can hold false beliefs, deceive others and pretend. Baron-Cohen argues (pg. 55) that a variety of other naturally developing axioms comprise our instinctual ToMM network, emphasizing, "children could also affirm a long list of axioms that constitute the core of their theory of mind, though as yet only a fraction of these have been explicitly stated and tested (such as 'seeing leads to knowing,' 'appearance is not necessarily the same as reality'….)." Baron-Cohen contends that, while children may not be explicitly aware of these ToMM 'theories,' they automatically put them to use in interpreting and

predicting the behaviors of others. He provides the following summary of ToMM, "it has the dual function of representing the set of epistemic mental states and turning all this mentalistic knowledge into a useful theory" (pg. 51).

Baron-Cohen (1997) proposes that the dramatic cognitive ToM deficits exhibited by autists stem primarily from a dysfunctional ToMM network, which is responsible for 'belief' processing. As opposed to normals and children with Down Syndrome, autistic children are typically unable to pass false-belief tests (the small minority that succeed do so at a later stage in development), which requires an understanding that other peoples' beliefs may differ from one's own (pg. 71). Accordingly, in picture sequencing studies, children with autism typically perform normally on tests involving physical causality and the attributions of basic desires and goals, but fail on tasks requiring an understanding of belief (pg. 71). Baron-Cohen argues that these experimental results are indicative of an "autistic-specific deficit in understanding beliefs as psychological causes of behavior" (pg., 72). Consistent with this thesis, additional autism studies have revealed a variety of other characteristic belief-related deficits, such as a failure to comprehend deception and engage in pretend play, distinguish between appearance and reality (i.e., realize that something may be different than it appears) and understand the concept of 'knowing' (Baron Cohen, 1997, Chapter 5). Baron-Cohen also reports that, although they can understand simple emotions such as happiness and sadness, autists struggle to identify more complex, belief-based emotions, such as surprise (pg. 78).

Baron-Cohen (1997) acknowledges that the four-part ToM system he postulates is highly speculative, and I will not offer a detailed evaluation here. Clearly, however, his skeletal account of affective ToM processing, focusing primarily on the Intentionality-

Detector, needs to be supplemented by a mirroring-based theory of the sort outlined in the previous section. Although he does not explicitly divide his ToM system along affective and cognitive dimensions, the indication is that the Intentionality-Detector is primarily an *affective* ToM network while the other three components (the Eye-Direction Detector, Shared-Attention Mechanism and Theory of Mind Mechanism) are more responsible for *cognitive* ToM processing. Accordingly, in describing ToM deficits in autism, Baron-Cohen emphasizes the characteristic failure to generate representations of belief, a *cognitive* ToM operation, as a core limitation. He speculates that this, in turn, may be the root cause of this group's inability to represent complex, "belief-based" emotions, such as shame and guilt; while underscoring that basic affective ToM processes involving the representation of simple desires (e.g., 'Bob wants the ball') and emotions (e.g., 'Sally is happy') seem to be spared.

Consistent with this diagnosis, there is mounting neurological evidence (see, for example, Blair & Blair, Shamay-Tsoory, and Pfeifer & Dapretto) that affective and cognitive ToM operations involve distinct, but interacting networks: a hypothesis receiving additional support from the psychopathy findings (i.e., psychopaths show more pronounced affective ToM deficits) described above. Addressing the neurological evidence, Shamay-Tsoory (2009) suggests, "the distinct abilities for cognitive and affective mental representation involve dissociable psychological and neural mechanisms and possibly engage discrete prefrontal circuitry" (pg. 221). Based on a series of lesion and fMRI imaging studies, Shamay-Tsoory contends that the cognitive ToM network—which includes the medial prefrontal cortex, superior temporal sulcus and the temporal poles—furnishes a capacity for self-other differentiation, third-person perspective taking,

and attributions of belief and intention. In contrast, the affective ToM system allows for an understanding of feeling and emotion. Shamay-Tsoory speculates that the mirroring system responsible for empathic state-matching provides inputs to the ventromedial cortex (VM); which, in turn, generates representations of affective states in coordination with the cognitive ToM network. In accord with this model, patients with VM damage primarily show affective, as opposed to cognitive, ToM deficits (pg. 223-224). Hence, neurological evidence supports a basic distinction between hot and cold ToM systems, in accord with the psychological evidence from psychopathy.

This finding has important implications for a contemporary philosophical debate concerning the neuropsychological underpinnings of our ToM capacities. In the philosophy literature, theorists have formed two main camps. "Theory theorists," such as Nichols & Stich (2003), argue that our capacity to understand the mental states of others depends on an implicit folk psychological theory, naturally interpreting the actions of others in terms of beliefs and desires. The ToMM device proposed by Baron-Cohen, which is hypothesized to generate mental representations of beliefs and intentions and axiomatic ToM rules, also appears to fit the theory-theory mold. Defining this general position, Shamay-Tsoory writes, "[theory theorists] maintain that mental states attributed to other people are conceived as unobservable, theoretical posits, invoked to explain and predict behavior, something akin to a scientific theory" (pg. 216). By contrast, "simulation theorists," such as Marc Iacoboni (2009), contend that ToM understanding relies on first-person processing, such that individuals internally simulate or recreate, as opposed to theoretically postulating, the mental states of others. Shamay-Tsoory characterizes this alternative view as follows, "one represents the mental states of others

by tracking or matching those states of one's own" (pg. 216). He also suggests that each philosophical position accords better with one of the two basic types of ToM processing, affective and cognitive, respectively. Affective ToM appears to incorporate relatively more simulation-based mechanisms, while cognitive ToM involves more theory-based operations (pg. 216). Accordingly, since it relies on similar mirroring mechanisms to those involved with affective ToM, emotional state-matching is also better suited to a simulation perspective. Addressing this point, Shamay-Tsoory writes, "with regard to the cognitive and emotional definitions of empathy, it may be suggested that cognitive empathy involves more [theory theory] processing, whereas affective empathy involves more simulation processing" (pg. 216). He, nonetheless, repeatedly emphasizes that the affective and cognitive ToM networks are constantly interacting in healthy individuals and "that a balanced activation of these two networks is required for appropriate social behavior" (pg. 228). Hence, it appears that while both theory-theory and simulation perspectives capture an important dimension of empathic processing, cognitive and affective, respectively, neither can stand alone. Given the wide range of processes involved, any adequate model of empathy must be a hybrid, like the one proposed by Alvin Goldman (2006), incorporating insights from both philosophical approaches.

As noted above, in his most recent book, Baron-Cohen (2011) emphasizes that autism characteristically involves *both* cognitive and emotional empathic deficits. With regard to the latter, he reports new findings indicating that, as compared to normals, autists show reduced sensorimotor response to pictures of people in pain (pg. 103), a diminished capacity to imitate emotional facial expressions (pg. 101), limited activity in 'emotional' brain regions (dorsomedial prefrontal cortex, posterior cingulate cortex, and

the temporal pole) during introspection exercises (pg. 102), and a decreased likelihood to interpret self-directed movement in terms of 'feelings' and 'desires' (pg. 101). Based on these findings regarding their emotional empathic deficits, and the overwhelming evidence that cognitive empathy is profoundly disrupted in this population, Baron-Cohen (2011) underscores that people with autism "show underactivity in almost every area of the empathy circuit" (100). The new findings regarding emotional empathic deficits in autists, however, are not necessarily inconsistent with the earlier research outlined by Baron-Cohen (1997) indicating that they possess some very basic emotional empathic capacities, such as an ability to recognize basic emotions. Similarly, the finding that they generally show diminished sensorimotor responsiveness to pictures of individuals in pain can also be squared with Blair & Blair's (2007) evidence that higher functioning autists retain some degree of autonomic sensitivity to distress cues (a response which is entirely lacking in psychopaths). Again, as noted above and repeatedly emphasized by Baron-Cohen (1997, 2011), autism is a spectrum disorder, encompassing a very wide range of neuropsychological and behavioral expressions—which makes it difficult to generalize about standard impairments. Nonetheless, based on current evidence, it appears that both cognitive and emotional empathy is characteristically diminished in this population, whereas psychopaths show a more selective deficit pertaining primarily to emotional empathy.

The typical moral shortcoming exhibited by autists, a failure to respond appropriately to the needs of others, can be readily accounted for by their diminished capacity for cognitive and emotional empathy. But why do autists typically not display the callous, violent behavior shown by psychopaths? As Baron-Cohen (2011)

emphasizes, unlike psychopaths, autists rarely *intend* to cause others harm (pg. 118). The explanation for this difference is not entirely clear. Blair & Blair (207) attribute this behavioral divergence to the emotional empathic capacities of autists, which Blair & Blair argue much surpasses that of psychopaths. Given doubts about emotional empathy in autism, however, it seems there must be more to the story. In general, autists are reluctant to engage with others and have great difficulty understanding motives and intentions, which would clearly limit their capacity for instrumental aggression. Indeed, it appears that psychopaths are so dangerous because their cognitive empathy abilities are largely intact in the absence of emotional empathic safeguards. This volatile neuropsychological profile allows them to intentionally manipulate, deceive, and harm others without any guilt.

Baron-Cohen (2011) offers an alternative explanation for the morally superior behavior of autists in comparison to psychopaths. He emphasizes that, unlike psychopaths, autists are capable of displaying deep ethical concern. Addressing this tendency, J. Kennett (2002) notes that individuals with Aspberger Syndrome typically approach morality in a Kantian way, viewing ethics norms as duties that are universally-binding for all people. Kennett writes, "autistic people…do seem capable of deep moral concerns. They are capable…of the subjective realization that other people's interests are reason-giving in the same way as one's own, though they may have great difficulty in determining what those interests are" (pg. 354). Baron-Cohen accounts for this proclivity on the basis of autists' penchant for "systematizing thought" and characteristic need for routine and order. He writes, "[Individuals with autism] are not like [psychopaths], for example, because though most people have developed their moral codes via empathy,

[autists] have developed their moral code through systematizing. They have a strong desire to live by rules and expect others to do the same for reasons of *fairness* (122). Baron-Cohen acknowledges that this moral orientation still leaves much to be desired. The same cold, drive for orderliness that furnishes their 'ethical' concern also leads autists to be inflexible in the application of the moral rules they construct and insensitive to the emotional repercussions. Nonetheless, as emphasized by Baron-Cohen, psychopaths appear incapable of displaying even this limited type of moral concern, a difference which may help to explain the divergent pattern of behavior exhibited by these two special populations.

This section focused on the distinctive empathic dysfunction characteristic of autism and the resulting moral profile. Unlike psychopaths, autists show profound cognitive empathic deficits in addition to emotional empathic limitations. As a result, people with autism have trouble identifying and responding to the needs of others; although they do not typically cause intentional harm to others and are capable of adopting a strict, rule-based approach to morality. Baron-Cohen (1997, 2011) indicates that although both are impaired in this population, affective and cognitive ToM processes rely on distinct, but interacting networks. The neurological evidence outlined by Shamay-Tsoory (2009) supports this view, as does psychopathy research indicating that cognitive empathy is largely intact in this population, despite their emotional empathic deficits. Emotional empathy, i.e., affective state-matching and related ToM processes, appears to be based on mirroring systems that, unless inhibited, automatically generate similar emotions to those perceived in others. By comparison to this simulation network, cognitive ToM processes (e.g., understanding belief and intention, advanced perspective

taking, etc.) seem to rely on relatively more detached, 'theory-based' operations. These findings may help to resolve the longstanding philosophical debate between "simulation theorists" and "theory theorists" concerning the neuropsychological underpinnings of our ToM capacities. It appears that both sides are partially correct, but any adequate model of the ToM system must incorporate both simulation and theory-theory perspectives to account for the distinctive mechanisms underlying affective and cognitive ToM processing, respectively. Indeed, although they are frequently interacting in healthy individuals, the evidence outlined above provides strong grounds for distinguishing between the more affective and cognitive dimensions of empathy—an approach that is consistent with the evolutionary evidence as well.

III. The Evolution of Empathy

In the *Age of Empathy*, Frans de Waal (2009) traces the evolutionary origins of empathy, drawing a similar distinction to the one endorsed here between its more cognitive and affective dimensions. He proposes a "Russian doll" model of empathy, emphasizing the evolutionary primacy of affective mirroring processes, as well as the important contributions made by cognitive ToM overlays. Characterizing this view, de Waal writes,

> Empathy engages brain areas that are more than a hundred million years old. The capacity arose long ago with motor mimicry and emotional contagion, after which evolution added layer after layer, until our ancestors not only felt what others felt, but understood what others might want or need. The full capacity seems put together like a Russian doll. At its core is an automated process shared with a multitude of species, surrounded by outer layers that fine tune its aim and reach. Not all species possess all layers: only a few take another's perspective, something we are masters at. But even the most sophisticated layers remain firmly tied to its primal core (pg. 208-209).

According to de Waal's evolutionary model, higher forms of sympathetic behavior are scaled up from more basic emotional processes. As noted in section I, de Waal endorses a PAM model of affective empathy proposing that the representation of another's mental state automatically activates a similar response in the observer, unless inhibited by prefrontal activity. He underscores that, based on this state-matching system, a great variety of species display personal distress responses to the perceived suffering of others. Studies have shown, for instance, that rats find the distress of conspecifics aversive, such that they will not take food if this results in another being shocked. Mole rats huddle closely together under stressful conditions. De Waal refers to this huddling behavior as an example of "preconcern," a rudimentary, more egocentric type of sympathetic behavior relying on very limited ToM capacities (pg., 95-96). He emphasizes, however, that the basic PAM mechanisms involved in these more automatic behavioral responses are also operative in more complex sympathetic responses. He proposes that, in general, every type of sympathetic behavior involves an affective

component (e.g., care or concern), which is funded by these PAM systems of emotional empathy.

In tracing the evolutionary roots of advanced sympathy, De Waal focuses primarily on two forms, "consolation" and "targeted helping." Among primates, such behaviors have only been observed in apes, our closer relatives. De Waal defines targeted helping "as assistance geared towards another's specific situation or need" (pg. 92). He has documented several examples in apes. For instance, he describes a case in which Kuni, a bonobo, picked up an injured bird in her enclosure, climbed with it to the top of a tree, spread its wings and flung it into the air in apparent attempt to help it fly. While these inter-species cases are quite rare—expressing a high level of perspective-taking--examples of helping and consolation within ape groups are ubiquitous. For instance, de Waal reports that apes have been observed adjusting their behavior to accommodate injured or disabled members within their group, for example, by being less rough in play or providing prolonged motherly care. Another type of targeted helping is consolation behavior, which de Waal characterizes "as reassurance by an uninvolved bystander to one of the combatants in a preceding aggressive incident" (2006, pg. 33). He reports that, in such cases, consoling parties more often sooth recipients of aggression versus instigators, and are also more apt to console victims of severe, as opposed to mild, aggression (2006, pg. 34-35)—which is consistent with a view that consolation behaviors are elicited by sympathy for those in distress.

De Waal emphasizes that these complex helping behaviors reflect core ToM capacities, such as an awareness of the distinction between self and other, as well as a basic understanding of the desires and goals of another. Addressing this point, de Waal

writes, "such cases illustrate the two-tiered process underlying helping: emotion and understanding. Only when both processes are combined can an organism move from preconcern to actual concern, including the targeted helping typical of our close relatives" (pg. 101). In accord with the view propounded by Decety and Lamm (see section I), de Waal speculates that self-other differentiation allows for the emergence of advanced forms of sympathetic behavior, as opposed to personal distress. According to de Waal's "co-emergence hypothesis," a capacity to display advanced sympathy, such as targeted helping, appears in conjunction with a sense of self (pg. 123), because both require similar cognitive ToM perspective-taking abilities. He speculates that the prefrontal processes involved in self-other differentiation or "mental separation" also allow for the inhibition of automatic emotional contagion. Addressing this connection and again emphasizing how the two main facets of empathic processing—affective and cognitive—work together, de Waal writes, "advanced empathy requires both mental mirroring and mental separation. The mirroring allows the sight of another person in a particular emotional state to induce a similar state in us….But we go beyond this, and this is where mental separation comes in. We parse our own state from the other's" (pg. 124). He notes that, as predicted by the co-emergence hypothesis, all the known species capable of passing mirror self-recognition tests, humans, apes, elephants and dolphins, also appear to display advanced forms of sympathy.

Although it would be beyond our scope to delve deeply into experimental research concerning the ToM abilities of chimpanzees, it is worth highlighting that chimps do appear to possess a basic understanding of goals and desires, as de Waal's

account of targeted helping suggests. In a recent review of the chimpanzee ToM literature, Call and Tomasello (2008) conclude,

> All of the evidence reviewed here suggest that chimpanzees understand…others in terms of a relatively coherent perception-goal psychology in which the other acts in a certain way because she perceives the world in a certain way and has certain goals of how she wants the world to be…But chimpanzees probably do not understand others in terms of a fully human-like belief-desire psychology in which they appreciate that others have mental representations of the world that drive their actions even when those do not correspond to reality (pg. 191).

In contrast to their successful performance on tasks relating to basic goal attribution, chimps perform poorly on experimental tests of false belief. Call and Tomasello report, "there is currently no experimental evidence that [chimps] understand false beliefs by, for example, predicting what another will do based on what the other knows (pg. 190). These findings, however, are at odds with reports of chimpanzee deception in more naturalistic settings, which, in turn, would seem to require some degree of false belief understanding. Hence, the negative experimental findings in chimps may be the result of methodological limitations, although a variety of nonverbal false belief tests have all yielded similar results (pg. 191). Regardless, chimp ToM research appears to support a distinction between two basic types of ToM orientations, roughly corresponding to the contrast drawn in the preceding section between the affective and cognitive ToM networks: a goal-based psychology capable of identifying basic aims and desires, versus a more cognitively flexible, belief-based psychology. Chimps appear to posses the former, but perhaps not the latter. Accordingly, while a goal-based psychology may

suffice for a range of advanced sympathetic responses, such as consolation and targeted helping, a belief-based psychology may be necessary for the broader range of sympathetic behaviors characteristically displayed by normally functioning adult humans.

Indeed, while generally emphasizing continuities with nonhuman primates, de Waal also reflects on the distinguishing characteristics of advanced forms of human sympathy. He suggests that what might set us apart is our capacity to expand the circle of our sympathetic concern, beyond immediate family and close associates. In general, primates, including humans, tend to empathize most readily with others who are geographically close, similar in physical appearance and with whom they have had repeated contact. This makes sense from an evolutionary perspective. Addressing this point, de Waal underscores, "empathy builds on proximity, similarity, and familiarity, which is entirely logical given that it evolved to promote in-group cooperation" (2009, pg. 221). He emphasizes that, as compared to our modern, 'globalized' society, our evolutionary ancestors lived in relative isolation, as members of small groups or bands. From a Darwinian perspective, empathy is designed to promote in-group cohesion, among individuals who are in frequent contact with one another. Accordingly, we seem naturally more inclined to empathize with individuals whose suffering we can directly witness. Addressing this issue, de Waal observes, "we care more about what we see firsthand…We're certainly capable of feeling for other based on hearing, reading, or thinking about them, but concern based purely on the imagination lacks strength and urgency" (pg. 221).

Despite these limitations, this capacity to empathize with distant individuals through imaginative identification may be unique to humans. Through reflection, we can

increase the size of our 'in-group,' acknowledging the rights and needs of other species, geographically remote individuals, and, sometimes, humanity as whole. Although such accomplishments are not as common as we would wish, this capacity for far-reaching empathic identification is impressive. Addressing this capacity, De Waal writes, "empathy's chief portal is identification. We're ready to share the feelings of someone we identify with, which is why we do so easily with those who belong to our inner circle. Outside the circle, things are optional" (pg. 213). In line with this theory, de Waal (1996) has proposed a "floating pyramid" model of altruistic behavior to explain under what conditions our parochial tendencies can be expanded. According to this theory, we are hardwired such that the range and extent of our sympathetic behavior is constrained by our health, security and level of material comfort. In general, altruistic acts directed at more remote (i.e. genetically, culturally, geographically) people is only possible--but certainly not guaranteed--when our basic survival needs are met. If resources are scarce, kin and close relations naturally take precedence. As this pressure is alleviated, sympathetic behaviors that are broader in scope can rise to the surface. Summarizing this view, de Waal writes,

> Altruism is bound by what one can afford. The circle of morality reaches out farther and farther only if health and survival of the innermost circles are secure. For this reason, rather than an expanding circle I prefer the image of a floating pyramid. The force lifting the pyramid out of the water—its buoyancy—is provided by the available resources. Its size above the surface reflects the extent of moral inclusion. The higher the pyramid rises the wider the network of aid and obligation (pg. 213).

Again, however rare, the fact that we are capable of overcoming our nepotistic leanings through imaginative identification is an impressive feat, facilitating some of our highest moral accomplishments—e.g., donating money to distant charities, affording equal rights to minority groups, etc. These examples demonstrate the dramatic role that advanced cognitive ToM processes can play in directing our natural sympathies outward; which again reinforces the ethical importance of having both facets of empathy, cognitive and emotional, function together.

IV. Moral Conscience: Lessons from Empathy

In comparison to the wide array of scientific writings on empathy, the neuropsychological foundations of moral conscience and the related emotions of shame and guilt have received much less attention, which is surprising given that discussions of empathy often address the topic of conscience. For instance, psychopaths, who demonstrate clear emotional empathic deficits, are also typically described as being "without conscience." Moreover, June Tagney and Ronda Darley (2002), two psychologists who have worked extensively on distinguishing shame from guilt at a psychological-behavioral level, emphasize, "guilt and empathy appear to work hand in hand in a mutually enhancing fashion" (pg. 89). Based on a variety of studies, they argue that shame and guilt are chiefly distinguished by their distinctive objects: self versus other. Unlike the more self-oriented emotion of shame, in pure guilt experiences the focus is on repairing a damaged relationship, righting a wrong done to another. Tagney & Darley explain,

> By its very nature, guilt forms a bridge to other-oriented empathic concern. In focusing on an offending behavior, the person experiencing guilt is relatively free of the egocentric self-involved process characteristic of shame. In fact, this focus on a specific behavior is likely to highlight the consequences of that behavior for a distressed other. In this way, guilt serves to foster an empathic connection (pg. 82).

Given the apparent connections, it seems plausible that empathy and conscience have similar neuropsychological dimensions. Indeed, conscience, and the related emotions of shame and guilt, appears to have both an affective and cognitive facet, very similar to that of empathy. For instance, with regard to the cognitive ToM component, it seems that in order to experience full-fledged guilt or shame one must, at minimum, be capable of differentiating between self and other and representing the beliefs and expectations of another. For this reason, autists typically neither experience shame and guilt nor understand these more belief-based emotions (Baron Cohen, 1997, 2011). On the other hand, it appears that psychopaths lack the requisite emotional repertoire to feel pangs of conscience. It seems that without the engagement of basic affective processes, such as responsiveness to anger, disappointment, distress-cues, etc., shame and guilt probably cannot emerge. Again, based on the empathy research cited above, it is likely that these affective responses are rooted in mirroring systems, which are partially separable from the prefrontal processes involved with the cognitive ToM aspects of conscience.

Accordingly, with regard to the evolution of conscience, one finds a pattern similar to empathy. It appears that human conscience is also constructed like a Russian

doll, with widely-shared affective processes at its core, and more evolutionarily recent cognitive ToM overlays. As cognitive ToM capacities grow, more advanced types of guilt and shame can emerge, culminating in highly reflective forms that only humans may be capable of displaying. Consistent with this approach, de Waal (1996) has suggested that basic norm internalization mechanisms, which are found in a variety of species, including dogs and primates, may be an evolutionary precursor to more advanced forms of guilt and shame. One prime example cited by de Waal is the 'cat and mouse' games waged between subordinate and alpha male macaques in the pursuit of valued mating partners. In the presence of alpha males, subordinates are typically deferential and careful not to pursue mates that are 'off limits.' As soon as alpha males are out of sight, however, subordinates readily sneak around and engage in illicit acts, with some awareness of the risks, as evinced by their nervous behavior (e.g., peeking inside a door to ensure that an alpha is at a safe distance). Moreover, subordinate violators subsequently show behavioral adjustment in the presence of alphas, displaying relatively higher levels of avoidance and submission after a transgression. Again, this behavior reflects awareness, although perhaps only a tacit one, of social norms and the possible consequences of violating them. Addressing the evolutionary significance of these behaviors, de Waal writes, "social rules among primates are not simply obeyed in their presence and forgotten in their absence…[and this may have] provided the starting point in the primate lineage for the evolution of a capacity for guilt and shame" (pg., 111). Further addressing the origins of guilt, he emphasizes, "anticipation of punishment and fear of endangering a valued relationship are not unrelated to guilt" (pg. 108). In a like vein, Tagney and Darley have argued that shame responses developed as signals of

appeasement by subordinates in hierarchical social settings, such as the ones found in chimp groups. These displays may have served an adaptive function by leading to withdrawal behavior that prevented escalation of tensions and conflict (i.e., punishment by dominants) after the violation of a socially-enforced norm (Tagney & Darley, pg. 126). Again, similar to the evolution of empathy, what appears to separate the forms of guilt and shame displayed by our primate relatives from the more advanced types found in humans is the cognitive ToM sophistication characteristic of the latter. Acknowledging these more cognitive facets of human conscience and shame and guilt experiences, de Waal writes, "we are dealing with complex emotions indeed. So complex, in fact, the term 'emotion' does not do them justice: self-consciousness, perspective-taking, and attribution are also involved" (pg. 109). The basic affective responses linked to these complex emotions, nonetheless, appear to have a long evolutionary lineage.

It would be beyond the scope of this chapter to offer a more detailed analysis of human conscience in all of its intricacies. I hope to have shown, however, that a similar approach to the one pursued here for the study of empathy—a method distinguishing between its cognitive and affective dimensions—may also prove valuable in related areas. Indeed, just like empathy, it appears that human conscience also has two faces.

Chapter 2 References

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, Mass.: MIT Press.

Baron-Cohen, S. (2001). Theory of Mind in Normal Development and Autism. *Prisme, 31:* 174-183.

Baron-Cohen, S. (2011). *The Science of Evil.* New York, NY: Basic Books.

Batson, C.D. (2009). These Things Called Empathy: Eight Related but Distinct Phenomena. In Decety, J., & Ickes, W. (eds.), *The Social Neuroscience of Empathy*. Cambridge, Mass.: MIT Press.

Blair, R.J.R., Mitchell, D., & Blair, K. (2005). *The Psychopath: Emotion and the Brain*. Malden: Blackwell.

Blair, R.J.R., & Blair, K. (2009). Empathy, Morality, and Social Convention: Evidence from the Study of Psychopathy and Other Psychiatric Disorders. In Decety, J., & Ickes, W. (eds.), *The Social Neuroscience of Empathy*. Cambridge, Mass.: MIT Press.

Call, J., & Tomasellom M. (2008). Does the Chimpanzee Have a Theory of Mind? 30 Years Later. *Trends in Cognitive Science,* 12: 187-192.

De Waal, F. (1996). *Good Natured: The Origins of Right and Wrong in Humans and Other Animals.* Cambridge, Massachusetts: Harvard UP.

De Waal, F. (2006). Morally Evolved: Primate Social Instincts, Human Morality, and the Rise and Fall of "Veneer Theory." In de Waal, F (ed.), *Primates and Philosophers* (pp. 1-58). Princeton: Princeton UP.

De Waal, F. (2009). *The Age of Empathy: Natures Lessons for a Kinder Society.* New York: Harmony Books.

Decety, J.,& Lamm, C. (2009) Empathy versus Personal Distress: Recent Evidence from Social Neuroscience. In Decety, J., & Ickes, W. (eds.), *The Social Neuroscience of Empathy*. Cambridge, Mass.: MIT Press.

Eisenberg, N., & Eggum, N. (2009). Empathic Responding: Sympathy and Personal Distress. In Decety, J., & Ickes, W. (eds.), *The Social Neuroscience of Empathy*. Cambridge, Mass.: MIT Press.

Goldman, A. (2006). *Simulating Minds*. Oxford: Oxford UP.

Hare, R. D. (1993). *Without conscience: The disturbing world of the psychopaths among us.* New York: Pocket Books.

Hume, D. (1978). *A Treatise on Human Nature*. 2nd Ed. Oxford: Oxford UP.

Iacoboni, M. (2008). *Mirroring People: The Science of Empathy and How We Connect with Others*. New York: Picador.

Kennett, J. (2002). Autism, Empathy and Moral Agency. *Philosophical Quarterly*, 52: 340-357.

Kennett, J. (2008). Reasons, Reverence and Value. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Kennett, J., & Fine, C. (2008). Internalism and the Evidence from Psychopaths and "Acquired Sociopaths." In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Kiehl, K. (2008). Without Morals: The Cognitive Neuroscience of Criminal Psychopaths. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

McGeer, V. (2008). Varieties of Moral Agency: Lessons from Autism (and Psychopathy). In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Nichols, S., & Stich, S. (2003). *Mindreading.* Oxford: Oxford UP.

Pfeifer, J., & Dapretto, M. (2009). "Mirror Mirror in My Mind:" Empathy, Interpersonal Competence, and the Mirror Neuron System. In Decety, J., & Ickes, W. (eds.), *The Social Neuroscience of Empathy*. Cambridge, Mass.: MIT Press.

Preston, S., & de Waal, F. (2002). Empathy: its ultimate and proximate bases. *Brain and Behavioral Sciences*, 25: 1-72.

Shamay-Tsoory, S.G. (2009). Empathic Processing: Its Cognitive and Affective Dimensions and Neuroanatomical Basis. In Decety, J., & Ickes, W. (eds.), *The Social Neuroscience of Empathy*. Cambridge, Mass.: MIT Press.

Tangney, J., & Dearing, R. (2002). *Shame and Guilt.* New York: Gulford.

Thagard, P. (2010). *The Brain and the Meaning of Life.* Princeton: Princeton UP.

3. Empirical Moral Psychology and the Future of the Weakness of Will Debate

In recent years, hope has been growing that vanguard research from a variety of scientific fields could finally help to settle longstanding metaethical disputes, especially those concerning our moral psychology. There are a great variety of metaethical issues. For instance, do moral judgments express beliefs or sentiments? Are there objective ethical truths? What is the meaning of moral concepts, such as 'good,' 'right,' 'just', etc.? Broadly conceived, ethical questions concern what we ought to do, while metaethical issues deal with a wider range of issues involving the nature of moral judgment, truth and justification. Based on this definition, one area of metaethics is descriptive ethics, which focuses on facts about our moral cognition. Citing the notorious is/ought problem, many moral philosophers remain sceptical about the *ethical* significance of empirical research. There has been less resistance, however, to empirical *metaethical* projects. Accordingly, Richard Joyce (2008) emphasizes that "the issue of whether a body of empirical data can have any metaethical implications is different from the issue of whether it can have ethical implications" (pg. 371). Indeed, I will argue below that contemporary research from neuroscience and psychology sheds new light on one of the oldest debates in metaethics: the so-called "weakness of will" problem.

The issue of what causes people to act immorally has been the subject of much controversy over the years. Are these failings due primarily to poor judgment or weakness of resolve? Historically, this debate focused on whether or not there are genuine occurrences of weakness of will or *akrasia*—i.e., cases where people know what they morally ought to do, yet fail to act accordingly. Some philosophers, such as Plato (1981)--at least in his early writings--seemed to deny this. On his view, we have a

motivational system that guarantees action in accord with moral belief: knowing "the good" suffices for doing "the good." Hence, immoral behavior always stems from a lack of moral understanding--false ethical beliefs concerning what we ought to do--rather than motivational weakness. Other ethicists, such as Augustine (1961) and Kant (1981), challenged the Platonic view, insisting that people may know what they ought to do, but nonetheless fall short. Passions or appetites can overwhelm our better judgment leading us to act in ways we know to be morally insufficient. Today, rejecting the platonic view, most moral philosophers agree that there are genuine occurrences of *akrasia.* There is still widespread debate, however, concerning the prevalence and scope of this phenomenon.

The internalism-externalism issue in metaethics is a modern outcropping of the ancient weakness of will debate. The *psychological* internalism-externalism problem, which will be carefully distinguished from other versions of this issue below, focuses on the relationship between moral judgment and motivation. The issue concerns whether we can make genuine[1] moral judgments without being motivated by them. If individuals judge that they ethically ought to perform action, Ω, does this mean that they will also be motivated to perform Ω? Internalists answer in the affirmative, while externalists deny that ethical judgment is automatically motivation-carrying. With regard to the historical weakness of will issue, psychological internalism appears to be more closely aligned with the Platonic view that ethical understanding, rather than motivational fortitude, is the key component for securing moral behavior. According to this line of thinking, if agents are naturally motivated by their judgments (i.e., psychological internalism is true), then

[1] As discussed in Section I, the issue of what constitutes "genuine" ethical assessment and how to identify if a population possesses this capacity is a subject of controversy.

making the right judgments in the first place is of central importance. Alternatively, in the tradition of Augustine and Kant, psychological externalists emphasize that making good judgments will often not suffice. Since some of our ethical assessments are motivationally inert, the pivotal moments for generating moral action occur after these judgments have been made--during the motivational phase of processing when strength of resolve comes heavily into play.

I propose below that empirical research supports a qualified version of psychological internalism (PI), holding that, in normal subjects, moral judgments carry motivational force. This inference is based on evidence suggesting that psychological noncognitivism (PN)—a thesis stipulating that ethical judgment is typically influenced by affect---is true. According to the prevailing view in psychological internalist-externalist literature, emotion is naturally motivation-carrying. Hence, granting this assumption, we can conclude on the basis of PN that PI holds as well. I argue, however, that some types of judgments carry greater motivational force than others. For instance, Joshua Greene distinguishes between two kinds of ethical assessment—"alarm bell" (AB) versus "somatic marker" (SM)--based on the differing types of emotional input linked to each. AB judgments are hypothesized to involve relatively more intense affect. Accordingly, I speculate that this type of judgment leaves less room for weakness of will than SM judgments, since the former class generally issues a stronger motivational pull.

While these findings do not resolve the contemporary weakness of will debate, they do provide some insight. Based on these findings, I conclude that *akrasia* occurs more frequently in the case of intuitive (i.e., rapid assessments in which conscious moral reasoning plays a minimal role), versus reason-based judgments, because the former class

characteristically involves more intense (AB) affect. Furthermore, I underscore that the connection between the psychological internalism debate and the controversy concerning the inner sources of immoral behavior is a more nuanced one than is commonly supposed. The finding that psychological internalism is true does not automatically imply that poor judgment is the root cause of unethical action, as some internalists have assumed. As defined here, moral motivation consists of *an inclination to act in accord with a moral consideration or judgment.* This definition allows for substantial variability across differing types of judgments, i.e.., with regard to their motivational force. In cases where motivational force is slight, unethical action could just as easily stem from weakness of resolve. Indeed, the key issue is not whether agents are morally motivated by their judgments, but rather if the motivational force will typically suffice for eliciting the action in question; in which case, we would be licensed to infer on the basis of PI that poor judgment, rather than weakness of resolve, is the primary source of immoral behavior. While the available evidence is limited, I believe it is highly unlikely that every psychological type of judgment—e.g., SM assessments--will typically carry motivational force sufficient for eliciting the respective action. Nonetheless, I think we can justifiably conclude that, as compared to cases involving reason-based assessments, unethical action following intuitive judgments will more often stem from poor judgment, as opposed to weakness of resolve. One consequence of this finding is that normative prescriptions focusing on strengthening resolve may be better suited to improving behavior that follows from reason-based judgments. In contrast, in the case of intuitive assessments, more emphasis should be placed on enhancing moral perception—making the right judgment in the first place.

I. Defining the Internalist Thesis

Before we can determine whether the internalist thesis has empirical support, we must first define its scope. This has been a matter of controversy within the internalist-externalist literature, with the leading proponents on either side disagreeing about what internalism means. Adina Roskies (2003), an avid critic of internalism, characterizes this position as holding "that motivation is intrinsic to, or a necessary component of, moral belief or judgment…If an agent believes that it is right to Ω in circumstances C, then he is motivated to Ω in C. " (pg. 52-55). She defines this as a metaphysical thesis that "purports to state a necessary truth about ethics" (pg. 52). On her view, in order to be "philosophically interesting," the internalist thesis must apply to any agent capable of making moral judgments. She criticizes alternative, *ceteris paribus* versions of internalism--for instance, one limiting its scope to particular types of agents (e.g., those that are 'normally functioning')—as being "philosophically anemic" and not in accord with the traditional view. Roskies writes, "the addition of the qualifying 'usually' or 'normally'…turns the internalist claim from a metaphysical one about an essential aspect of moral judgment to a merely descriptive claim about what is generally the case" (pg. 54). According to Roskies' strict definition, just one counterexample to the internalist thesis—i.e., a genuine moral judgment that does not carry motivational force--would undermine this position. As it happens, Roskies argues that VM patients are "walking counterexamples to the strong internalist claim" (pg. 55). Elsewhere, she suggests that psychopaths may also refute internalism, writing, "I question whether only people with unimpaired judgment are potential counterexamples" (2008, pg. 201).

On the other side of the internalist-externalist debate, Jeannette Kennett and Cordelia Fine (2008), proponents of internalism, believe that Roskies definition is overly restrictive. Kennett and Fine argue that "internalists claim [only that] in one way or another [genuine moral judgment and moral motivation] are internally connected" (pg. 173). *Contra* Roskies, Kennett and Fine suggest that *ceteris paribus* versions of internalism, which avoid positing a metaphysically necessary connection between judgment and motivation, are philosophically interesting and more empirically plausible. For instance, they note, "a brand of internalism arising from evolutionary accounts of morality will restrict its claim to normally functioning individuals….we see no reason in principle why its proponents couldn't allow that certain conditions such as depression might block the motivation that normally flows from moral judgment (pg. 180)." While they stop short of endorsing any particular version of *ceteris paribus* internalism, Kennett and Fine disagree with Roskies' claim that evidence from clinical populations undermines this general position. They insist that, on the contrary, this research "consistently supports an association between deficient moral behavior and deficient moral understanding…[which] tends to support rather than undermine the general thrust of [internalist] claims" (pg. 189).

The philosophical question regarding what constitutes an "interesting" version of internalism cannot be resolved by empirical means alone. Scientific research may be able to provide evidence that either supports or fails to support an internalist thesis, but it cannot tell us whether this thesis is philosophically meaningful. As a pragmatist, I believe that the value of a philosophical thesis is determined by the concrete problems it can address and, ideally, improve. Hence, while questions of philosophical significance

are not empirically reducible, these issues are nonetheless anchored to concrete realities on the ground. I must confess that, given this orientation, I find any talk, including Roskies', of a metaphysical thesis dealing with "necessary" truths problematic. What exactly is a metaphysically necessary truth, and how would this fit into a naturalistic worldview? Setting this more philosophical concern aside, I believe that Roskies' traditional approach to the psychological internalist-externalist issue should be rejected on pragmatic grounds: it has led to a stalemate that can only be overcome by a fundamental change in methodology. Viewing the internalist thesis as an all-or-nothing proposition that must apply to every kind of moral personality, contemporary researchers, like Roskies, have focused on identifying just one counterexample, believing this should suffice for rejecting internalism *tout court*. As a result, the current debate in the internalist-externalist literature focuses on whether special populations, such as VM patients or psychopaths, are "walking counterexamples" to the internalist thesis. This approach has led to an impasse because researchers disagree about whether these special populations have the capacity for genuine moral judgment. As I will argue below, the verdict is still out, and I do not believe that a resolution will be reached anytime soon—which provides good reason to adopt another, more promising approach to the internalism question.

Before proceeding, I must clarify what version of the internalist-externalist debate is at issue here. The general debate concerns the motivational force of moral judgments. There are several dimensions to this question, however, each of which has been the subject of controversy. Roskies, Joyce and Kennett and Fine focus primarily on the *psychological* question of whether genuine moral judgments are naturally motivating. In

other words, are we automatically motivated to act in accord with our ethical assessments? Psychological internalists answer in the affirmative, while psychological externalists, like Roskies, insist that moral judgment and motivation are cognitively dissociable. The psychological internalist-externalist debate must be carefully separated from the conceptual version of this issue. The latter primarily concerns the *meaning* of moral terms. Conceptual internalists posit that moral judgment conceptually entails motivation: 'being motivating' is part of what the concept 'moral judgment' means. For instance, Michael Smith (2008) argues in favor of a rationalist version of conceptual internalism. Smith posits that, according to the lay conception, agents are either motivated by their ethical judgments *or they are irrational*.

Shaun Nichols (2008) proposes that questions about the meaning of moral terms, such as the conceptual internalist thesis, can be resolved by surveying folk intuitions. This conviction is based on an assumption that I will not question here--that the meaning of ethical terms is determined largely by the way they are commonly used and conceived. Seeking to empirically test conceptual internalism, Nichols (2002a) conducted a study of folk views regarding the connection between moral judgment and motivation. He discovered that, when asked about special agents—such as psychopaths or the devil—people tend to claim that these agents understand the difference between right and wrong, yet are unmotivated by their judgments. Nichols contends that these findings suggest that people believe it is possible to make genuine moral assessments without being motivated by them, and hence the folk concept of ethical judgment appears to be externalist in character.

Assuming Nichols' interpretation is correct, does this conceptual finding entail that externalism as a *psychological* thesis is true? Clearly, it does not. The psychological and conceptual internalist-externalist debates are separate empirical questions, for which different kinds of evidence are relevant. Emphasizing this divide, Kennett and Fine write, "we do think there is a conceptual connection between moral judgment and motivation, but we do not think that the connection *is thereby [psychologically] guaranteed*" (pg. 217). Drawing a similar distinction between Smith's conceptual rationalism and psychological rationalism (i.e., a thesis stipulating that our moral judgments are psychologically produced via reasoning), Nichols (2004a) underscores, "rationalists thus make both the conceptual and [psychological] claims, and what I want to stress for present purposes is that the claims are independent. Either of the claims could be true while the other claim is false" (pg. 69). The conceptual internalism-externalism debate concerns our *beliefs about* the psychological underpinnings of ethical judgment. Nichols' argues that, according to the folk view, moral judgment does not psychologically entail ethical motivation. Of course, the folk conception could be wrong. One important lesson from modern cognitive science is that introspection is not always a reliable guide to our inner cognitive workings. Our common psychological conceptions may not match the underlying reality. Indeed, I will argue below that neurocognitive evidence supports psychological internalism, and hence the externalist folk view, assuming the folk are externalists as Nichols' contends, is false.

As noted above, the current literature on the psychological internalist-externalist issue focuses on research with special populations. One population that has received a lot of attention in connection with this question is psychopaths. Do these individuals know

what is morally right and wrong, yet remain unmotivated by this awareness, or do their well-documented emotional/empathetic deficits (see Chapter 2) short-circuit the capacity for genuine moral understanding? Is the source of their aberrant behavior weakness of will or defective judgment? It seems that psychopaths may be, to borrow Roskies apt phrase, "walking counterexamples" to the internalist thesis—a group of people who make genuine moral judgments, but feel no motivation to act in accord. Marc Hauser's theory (2006), as described in Chapter 1, is consistent with this view. In developing his Chomskian account of intuitive moral judgment, Hauser postulates that psychopaths are generally morally competent, but lack the affectively-based motivational system to follow through. Other theorists, such as Shaun Nichols (2004a), speculate that these affective deficits diminish psychopaths' capacity for genuine ethical judgment, and so they cannot serve as counterexamples to internalism.

More recently, the focus of the internalist-externalist literature has shifted from psychopaths to VM patients, another clinical group with affective deficits. While members of this population generally do not display the extreme antisocial behavior characteristic of psychopaths, VM patients nonetheless act in ways—e.g., neglecting familial responsibilities, taking risky gambles that put others in jeopardy, failing to honor promises and commitments, etc.--that can plausibly be deemed immoral. As touched upon above, Roskies argues that members of this special population, people who appear to have normal declarative moral knowledge, yet fail to behave in a way consistent with this understanding---provide a counterexample to the internalist thesis. Outlining her diagnosis, Roskies' writes, "I believe the [ventromedial cortex] forms a causal connection from the cognitive to the affective systems. If this link is severed, one would anticipate

seeing judgment preserved but affect and motivation impaired, which is precisely the VM patient's clinical syndrome" (2008, pg. 192). Kennett and Fine take issue with several aspects of Roskies' argument. Primarily, they disagree with her contention that there is compelling evidence that VM patients are morally competent, without substantial deficits in their capacity for genuine ethical judgment.

Hence, regardless of whether it is the abilities of psychopaths or VM patients that is at issue, the general pattern in the internalist-externalist literature remains the same: one group of theorists argue that a particular clinical population undermines the internalist thesis, while another group of theorists claim that this clinical population is irrelevant to the internalist-externalist debate because this population lacks the capacity for genuine moral judgment. Addressing this issue, Kennett and Fine write, "the debate in moral philosophy has largely turned on whether the amoralist really makes moral judgments or only does so in an 'inverted commas' sense: that is, a sense that 'alludes to the value judgments of others without itself expressing such a judgment" (pg. 148). Kennett and Fine believe that the major problem with the prevailing empirical approach to the internalist-externalist issue is that both sides disagree about what constitutes genuine moral judgment, and, as a result, are merely talking past each other. Kennett and Fine emphasize,

> any attempt to settle the debate between internalists and externalists by empirical examination of putative counterexamples to internalism cases appears doomed while each side uses the term 'moral judgment' in a different sense…the psychopath might qualify as making a moral judgment in the externalist story of what this involves but fail to satisfy…internalist criteria…It should come as no surprise that the internalist claim turns out to be false when it is attached to an externalist account of moral judgment, and vice-a-versa (pg. 218-219).

In order to move beyond this impasse, Kennett and Fine propose two theory-neutral (i.e., neutral with respect to the internalist-externalist debate) criteria for determining whether a population possesses the capacity for genuine moral judgment: measures which apparently exclude both psychopaths and VM patients.

Before assessing their proposed solution, I want to focus in greater depth on Kennett and Fine's claim that internalist and externalist researchers are working with differing models of genuine moral judgment. Richard Joyce (2008) brings greater clarity to this problem by highlighting the connection between the internalism-externalism dispute and another longstanding metaethical debate, the cognitivist-noncognitivist issue. Specifically, he emphasizes that internalist researchers generally operate with a noncognitivist conception of genuine moral judgment, while externalists typically endorse a competing cognitivist model. As I will highlight below, Joyce fails to clearly distinguish between differing versions (i.e., psychological versus conceptual) of both the internalist-externalist and cognitivist-noncognitivist debates, rendering some of his conclusions erroneous. He is right, however, that the debate between internalists and

externalists—at least the psychological version--really boils down to a disagreement over which model of genuine moral judgment, cognitivist or noncognitivist, is correct.

As Joyce characterizes it, the traditional cognitivist-noncognitivist debate concerns the function of public moral utterances. Cognitivists claim that moral judgments express beliefs, and thus may be true or false, while noncognitivists deny this. Traditionally, the most popular form of noncognitivism has been emotivism. Joyce writes, "[emotivists claim that] when we make a moral judgment, we are not expressing a belief (i.e., are not making an assertion), but rather are expressing some kind of conative mental state, such as a desire, emotion or preference" (pg. 373). He further emphasizes that historically internalism and noncognitivism have been viewed as mutually reinforcing positions, just like externalism and cognitivism. "The biggest fans of motivation internalism in metaethics have traditionally been the noncognitivists..,[In contrast] most moral philosophers who embrace pure cognitivism see motivation internalism as an unlikely and unnecessary thesis" (pg. 387). Joyce explains this pattern by pointing out that philosophers have generally assumed that beliefs and desires are different types of mental states, and only desires are intrinsically motivating. Hence, noncognitivists naturally assume that internalism holds, since, according to their model, genuine moral judgments are desire-expressing. On the other hand, cognitivists, like Marc Hauser and Adina Roskies, suppose that genuine moral assessments are expressions of belief, which may not elicit corresponding desires. Hence, cognitivists are typically externalists.

In offering his diagnosis, Joyce uncovers a circularity problem in the prevailing approach to the psychological internalist-externalist issue. The current debate focuses on

whether psychopaths and VM patients are counterexamples to internalist thesis; but they can only serve this function if they make genuine moral judgments. In turn, whether or not these special populations possess this capacity will depend on one's model of genuine moral judgment. Both psychological internalists and externalists acknowledge that psychopaths and VM patients have affective deficits, but they disagree about the implications. Believing that affect is necessary for issuing genuine moral judgments, internalists will likely deny that these groups possess this ability. On the other hand, externalists, who are generally cognitivists, reach the opposite conclusion. Figure 2 below illustrates the circularity problem highlighted by Joyce:

Figure 2: The Circularity Problem

	Cognitivism-Noncognitivism	**Psychopaths Relevant?**	**VM Patients Relevant?**
Internalists	Noncognitivism	No	No
Externalists	Cognitivism	Yes	Yes

Joyce underscores that we cannot expect any further empirical progress on the internalist-externalist issue until researchers on both side can agree on a model, either cognitivist or noncognitivist, of genuine moral judgment. He emphasizes,

> if we treat moral judgment as a kind of linguistic performance [cognitivism]—as a speech act—then it is indeed reasonable to assume that these patients are capable of making moral judgments…If, on the other hand, we prefer to treat moral judgment as more of a psychological event [noncognitivism], as a kind of internal "mental assent" to an evaluative proposition then serious doubt arises…The notion of a moral judgment is sufficiently pliable to allow reasonable precisifications according to which internalism is pretty obviously false, and equally reasonable precisifications according to which it may be true (pg. 385-8).

Jesse Prinz (2006), a noncognitivist, commits an egregious example of this biasing error. In trying to bolster the case for internalism, Prinz asks us to consider whether someone can genuinely attest that "killing is wrong" without any corresponding sentiment. Prinz believes the answer is 'no,' and that this provides *prima facie* support for psychological internalism. Prinz's conclusion about the test case is based on his antecedent view of moral judgment: because he already endorses noncognitivism, Prinz has the intuition that his test subject fails to make a genuine moral judgment. Joyce's point is that a cognitivist will likely have a contrasting intuition.

Based on his analysis, Joyce concludes that neuroscientific research is irrelevant to the internalist-externalist issue. On his view, the only hope of resolving this issue is making progress on the cognitivist-noncognitivist debate, which he characterizes as a problem of conceptual meaning—i.e., concerning the function or meaning of public moral utterances. Accordingly, he suggests that socio-linguistic research, presumably of the sort practiced by "experimental philosophers," such as Shaun Nichols, will provide the only relevant empirical data for this enterprise. He summarizes his view as follows,

> the [truth of motivation internalism] depends on whether the sincere acceptance of a moral judgment implicates motivational structure, which in turn depends on whether there exists linguistic conventions according to which public moral judgments function to express…conative attitudes. To the extent that this is an empirical matter, it is a job for sociolinguistics; I see no obvious place for a contribution from neuroscience (2008, pg. 389).

Joyce's basic argument is that the cognitivist-noncognitivist dispute underlies the internalism/externalism debate; and since the former is primarily a socio-linguistic issue, we cannot expect neuroscience to shed light on the related internalism/externalism question.

In making this argument, Joyce fails to distinguish between two versions of the cognitivist-noncognitivist debate. As noted above, the traditional cognitivist-noncognitivist issue centers around the function of public moral utterances: are moral judgments *expressions* of beliefs or desires? Joyce builds his argument around this version of the cognitivist-noncognitivist debate, which closely parallels the *conceptual* internalist-externalist issue. Importantly, another version of the cognitivist-noncognitivist debate concerns the *psychological* question of what primarily causes our moral judgments, beliefs or desires? Psychological noncognitivists, such as Jonathan Haidt (2001) and Shaun Nichols (2004a), argue that affect largely determines the judgments we make. In contrast, psychological cognitivists, like Marc Hauser, propose that at least some of our ethical assessments (e.g., intuitive judgments) are driven by the cold processing of implicit beliefs. Importantly, these two versions of the cognitivist-noncognitivist question, the linguistic and the psychological, are separable. It is possible

that our ethical assessments may function linguistically as expressions of our desires (linguistic noncognitivism), but these desires could be caused by beliefs (psychological cognitivism), and vice-a-versa (linguistic cognitivism and psychological noncognitivism). In fact, Joyce acknowledges this disconnect, writing, "it is entirely possible that moral judgments are typically caused by emotional activity but nevertheless function linguistically as assertions (i.e., expressions of belief)" (pg. 375).

Hence, the vital question for our purposes is which version of the cognitivist-noncognitivist issue is germane to the psychological internalist-externalist debate? The only cognitivism-noncognitivism issue that seems directly relevant is, not surprisingly, the psychological question of what causes our moral judgments. Granting the general presumption that emotion is connected to motivation (more on this in Section II), if neuroscientific research provides evidence that our moral judgments are caused by affect (psychological noncognitivism), this would seem to bolster the psychological internalist position. Conversely, if our judgments are caused primarily by beliefs (psychological cognitivism), which are *ex hypothesi* not intrinsically motivating, then it seems plausible that we could make genuine moral judgments without being motivated by them. While Joyce may be right that the *linguistic* cognitivist-noncognitivist debate --which would seem to bear more directly on the *conceptual* internalism-externalism question--is primarily a socio-linguistic matter, the same cannot be said for the psychological issue. Cognitive and neuroscientific research is clearly relevant to the question of what causes our moral judgments--and, by extension, to the psychological internalist-externalist issue. Indeed, I will argue in the next section that evidence in favor of psychological noncognitivism provides the strongest support for psychological internalism.

While Joyce conflates differing versions of the cognitivist-noncognitivist and internalist-externalist debates, the circularity problem he highlights still remains. Given their opposing models of genuine moral judgment, psychological internalists and externalists disagree about what, if any, special populations are counterexamples to the internalist thesis. Internalists, who are typically psychological noncognitivists, believe that genuine moral judgment requires the right kind of sentiment, and thus they disqualify psychopaths and VM patients from consideration. Externalists, who generally endorse psychological cognitivism, hold a contrary view. For instance, on Hauser's account, moral judgment typically precedes and causes an affective-motivational response, and hence genuine ethical assessments can occur in the absence of sentiment.

In an effort to resolve this circularity problem, Kennett and Fine (2008) propose two theory-neutral criteria (i.e., neutral with respect to the psychological cognitivist-noncognitivist issue) for determining whether a population has the capacity for genuine moral judgment. One criterion would test if subjects draw the normal distinction between moral and conventional normative violations, while the other criterion would focus on subjects' ability to apply moral terms to a wide range of cases in a consistent way. The first test relates to the robust empirical findings (see Chapter 4 for more details) showing that normal subjects tend to judge moral violations (e.g., the unprovoked hitting of another person) as more serious, universal and less permissible than conventional violations (e.g., breaking rules of etiquette). The second, concept-application criterion is drawn from research on psychopaths, who tend to use moral terms in bizarre and contradictory ways. For instance, Kennett and Fine provide the following examples drawn from Hare (1993):

- An inmate described his murder victim as having benefitted from the crime by learning "a hard lesson about life."
- When asked if he experienced *remorse* over a murder he'd committed, one young inmate told us, "Yeah, sure, I feel remorse." Pressed further, he said that he didn't "feel bad inside about it."
- "My mother is a great person….I really *care* for that woman, and I'm going to make it easier for." When asked about the money he had stolen from her he replied, "I've still got some of it stashed away, and when I get out it's party time!"

Based on these samples, it seems doubtful that psychopaths are using moral terms, such as 'remorse', 'care', etc., in standard ways. Kennett and Fine write, "the erratic, inconsistent, and contradictory nature of their pronouncements suggests that [psychopaths] do not possess the moral concepts of nonpsychopathic individuals. They are incompetent in the use of evaluative terms" (pg. 176).

Before evaluating Kennett and Fine's two-pronged measure of the capacity for genuine moral judgment, it is worth considering whether psychopaths and VM patients pass the test. Psychopaths appear to fall short on both measures. Kennett and Fine offer the concept-application standard specifically with this group in mind, to disqualify any clinical population showing similar aberrations in the usage of moral language. As noted in the previous chapter, research also reveals that members of this population fail to draw the standard distinction between moral and conventional normative violations. Based on these measures, Kennett and Fine conclude that psychopaths do not have the capacity for genuine moral judgment, writing, "we claim that a growing body of evidence…such as their poor performance on the moral-conventional distinction task and their incompetence in the use of moral language, suggests that psychopaths…do not have mastery of the relevant moral concepts" (pg. 219).

In contrast, it is unclear whether VM patients pass both tests. The prevailing view is that, despite their affective deficits, individuals from this population possess intact moral knowledge, similar to that of normal subjects (see Damasio, 2000). There is some evidence in favor of this view. For instance, as reported by Roskies, a study by Young et al. (2006) found that VM patients' judgments of moral culpability and intention match those of normal subjects. A strong case could also be made that this group satisfies the concept-application standard, as VM patients do not show the same aberrations in the usage of moral terms characteristic of psychopaths. Summarizing the received view, Roskies writes, "there is a growing body of evidence attesting to the ability of VM patients to make [genuine] moral judgments" (2008, pg. 196). On the other hand, there is also evidence that VM patients' capacity for 'normal' moral judgment may be damaged. For instance, as described in Chapter 1, studies have shown that these subjects offer abnormal responses to the well-known trolley problem scenarios, tending towards consequentialist responses. For example, these patients judge that it would be permissible to push a bystander in front of an oncoming train to save five lives. In contrast, among normal subjects, approximately eighty percent claim this action would be impermissible. With regards to the moral/conventional standard, the only direct evidence comes from a study Blair and Cipolotti (2000) conducted with one VM patient, JS. Somewhat surprisingly, given the prevailing view that VM patients are morally competent, JS failed to distinguish between moral and conventional violations, in a way similar to psychopaths. Clearly, more studies need to be conducted with larger sample sizes before we can comfortably assert that VM patients fail the moral/conventional test. This finding from JS is suggestive, however. Hence, while VM patients likely satisfy the

concept-application standard, the same cannot be said regarding the moral/conventional measure.

Based on Kennett and Fine's test, both psychopaths and VM patients appear to lack the capacity for genuine moral judgment, and hence cannot serve as counterexamples to the internalist thesis. Kennett and Fine claim to have provided two criteria that are neutral between psychological cognitivism and noncognitivism. In regard to theory-neutrality, both of their measures appear to succeed. A deeper question concerns whether Kennett and Fine's test is a reliable indicator: is having the ability to distinguish between moral/conventional normative violations and use moral language competently indicative of a capacity for genuine moral judgment? It seems that Kennett and Fine have provided only two necessary--but not sufficient—conditions; and thus the circularity problems highlighted by Joyce still hamper the conventional approach to the psychological internalist-externalist issue.

Indeed, although Kennett and Fine's criteria resolve the psychopath and VM issue, it does not directly address the key dispute between psychological internalists and externalists—i.e., whether or not emotion is necessary for genuine moral judgment. Internalists, who are typically noncognitivists, answer in the affirmative, while externalists, who are generally cognitivists, deny this. As a result, psychological internalists and externalists could concur that other special populations would pass both of Kennett and Fine's tests, while still disagreeing about whether *these* populations possess a capacity for genuine moral judgment. Consider, for example, individuals with severe depression or high-functioning autists, two groups that also appear to suffer from distinctive moral impairments (e.g.., neglecting others needs' due to self-focus). It seems

plausible that both of these groups could pass Kennett and Fine's test. Do these groups possess a capacity for genuine moral judgment? Given that severe depression is linked to a general blunting of emotion, internalists would likely deny that individuals suffering from this disorder possess this ability; while externalists may reach a different conclusion. In contrast, given the more cognitive, as opposed to emotional, impairments characteristic of high-functioning autists (see previous chapter), externalists would probably be less inclined than internalists to attribute a capacity for genuine moral judgment to this group. Kennett and Fine's criteria cannot resolve the issue of whether these special populations are "walking counterexamples" to the internalist thesis, since both groups pass the test that disqualifies psychopaths and VM patients from consideration. Hence, Kennett and Fine have not provided sufficient resources for rescuing the conventional approach to the psychological internalist-externalist issue from problems of theory-laden interpretation. Once the attention shifts to special groups that pass Kennett and Fine's test, internalists and externalists will continue to argue about what *additional* capacities are necessary for genuine moral judgment.

In my opinion, however, there is an even deeper problem with the conventional approach to the psychological internalist-externalist question. It runs counter to a growing trend in moral psychological research attesting to the diversity of ethical impairments. In recent years, great strides have been made towards understanding the idiosyncratic mechanisms underlying the different deficits of distinctive groups, such as psychopaths, VM patients, people with autism, etc. As we continue to learn more about each type of moral deficiency, the tendency to group everyone together for the purpose of evaluating descriptive metaethical theses appears increasingly outmoded. The search for

"walking counterexamples" to the internalist thesis is based on a view that moral psychological theses must apply to everyone or else they are bankrupt. I believe this is no longer a sustainable approach. Whether internalism applies to unique groups is a worthwhile question to pursue--but we should not assume that the results from one sub-population can or should be generalized to all the rest.

Indeed, I will argue below that there is empirical support for the following *ceteris paribus* version of internalism:

Ceteris Paribus Internalism (CPI): *in normally functioning subjects, judging that action, Ω, is morally required typically carries with it a motivation to Ω.*

CPI is limited to "normally functioning" individuals, broadly construed, by which I mean people who do not suffer from brain impairments or mental illness, such that they would fall within a clinical population. Importantly, psychopaths and VM patients (as well as autists and severe depressives), the two groups that have received the most attention in the internalist-externalist literature, both fall outside of CPI's restriction to normally functioning individuals. By limiting our focus in this way, we can temporarily sidestep the issue of what constitutes a capacity for genuine moral judgments. Whatever this ability amounts to, all researchers would presumably agree that it is possessed by average members of the population. As a psychological thesis, CPI also avoids positing a metaphysically necessary connection between moral judgment and motivation, unlike Roskie's version of internalism.

Another error characteristic of the prevailing all-or-nothing approach to the internalism question closely parallels the one cited above. Metaphysically-minded theorists have posited that in order to be true the internalist thesis must not only apply to

everyone, but to every token judgment as well. According to this view, if just one example of a motivationally inert ethical assessment can be found, then internalism can be rejected *tout court*. In contrast, CPI posits that in normally functioning individuals there is a psychological connection between moral judgment and motivation, such that the latter will *typically* follow from the former, in the vast majority of cases. CPI allows that special circumstances—e.g., altered states resulting from drug use, highly stressful situations, etc.—may interrupt this natural link. Hence, CPI does not require that *every* ethical assessment made by normally functioning individuals must carry motivational force, but only those issued under standard conditions. In my view, requiring that the psychological connection between ethical judgment and motivation be a 'necessary' one—assuming it makes sense to speak in such terms, which is not at all clear to me—sets the bar too high. With regard to the psychological internalism debate, a *ceteris paribus* version of internalism seems like the only empirically plausible candidate.

As noted above, without providing a detailed explanation, Roskies claims that *ceteris paribus* versions of internalism are "not philosophically interesting." Clearly, I disagree. It would be beyond our scope to define what constitutes a philosophically interesting thesis, but I will point to one reason why CPI meets this standard. As I intend to show, the finding that this proposition is likely true has important implications for the contemporary weakness of will issue, a problem that has traditionally been of great interest to moral philosophers.

II. Psychological Classes of Judgment

In the preceding section, I argued that the empirical study of the psychological internalist-externalist issue has been hampered by a misguided tendency to treat the internalist thesis as an all-or-nothing proposition that must apply to everyone and every token judgment. As a result, researchers working in this area have focused primarily on whether psychopaths and VM patients can serve as counterexamples to this position, without reaching any consensus. One reason for the lack of progress is that psychological cognitivists and noncognitivists disagree about what counts as genuine moral judgment. Kennett and Fine's theory-neutral criteria-- which provide two *necessary*, but not *sufficient* conditions--still leave room for debate. Based on the manifest failure of this conventional approach, I suggested that it makes more sense to qualify the internalist thesis by limiting it to "normally functioning" individuals. Indeed, I proposed that we temporarily set aside questions concerning the moral judgment capacities of VM patients and psychopaths, and focus instead on empirically evaluating CPI, which posits, *in normally functioning subjects, judging that Ω is morally required typically carries with it a motivation to Ω.* While lacking the scope of traditional internalist theses of the sort challanged by Roskies, CPI nonetheless makes an important claim about the ethical judgments of a majority of the population. As I will argue below, this thesis has significant implications for the contemporary weakness of will debate.

Before evaluating CPI, one central challenge must be addressed. Theorists, like me, assessing theses concerning the nature of moral judgment must carefully avoid imputing too much uniformity. It is risky to make grand pronouncements about human ethical judgment in general when there appears to be such a rich variety. Many

contemporary researchers are sensitive to this issue. For instance, in considering the psychological internalist thesis, Kennett and Fine (2008) offer the following taxonomy of differing kinds of ethical judgments:

1. *Third Personal:* what someone should do.
2. *Second Personal:* what you should do (face-to face advice).
3. *First Personal:* what I should do.
4. *Armchair*: about hypothetical situations or about what kinds of principles we should adopt to govern our choices.
5. *In situ:* what should be done in these actual circumstances.

Kennett and Fine note that the internalist thesis could theoretically apply to some of these categories but not others. This is just one dimension from which to distinguish differing types of ethical judgment, and there are certainly many others. Hence, it is dangerous to make general claims about moral judgment *writ large.* Despite these concerns, I believe there is enough available evidence to reach a provisional verdict regarding CPI. As I will argue below, based on the reasonable assumption that psychological noncognitivism—a thesis that emotion causally influences the moral judgments issued by normal subjects--generally holds, there is reason to believe that CPI is true as well.

One prominent theorist who endorses psychological noncognitivism is Joshua Greene (2008). As described in Chapter 1, based on his FMRI research, Greene provides a "dual processing" theory of moral judgment that distinguishes between two main types—deontological versus utilitarian-style--each connected to a different type of emotional response. According to this theory, when we reach characteristically deontological conclusions (e.g., it is wrong to sacrifice one innocent life to save others), our judgments are intuitive and emotionally-driven, based on what Greene refers to as "alarm bell" (AB) affect. In contrast, characteristically utilitarian judgments (e.g.,

maximize the good, even if this requires the loss of innocent life) are relatively slower and guided by conscious moral reasoning, which is more subtly influenced by what Antonio Damasio (2000) refers to as "somatic markers" (SM affect). Greene emphasizes that both psychological types of judgment involve affective input, but of a different kind. Propounding a version of psychological noncognitivism, he writes,

> I am sympathetic to Hume's claim that *all* moral judgment (including consequentialist judgments) must have some emotional component. But I suspect that the kind of emotion that is essential to consequentialism is fundamentally different from the kind that is essential to deontology, the former functioning more like a currency and the latter functioning more like an alarm (pg. 41, emphasis added).

Before assessing Greene's basic taxonomy, it is worth considering how both types of affective input—AB versus SM--are supposed to shape our judgments in unique ways. As quoted above, Greene suggests that the former type of input functions like an "alarm" while the latter serves as "currency," but what exactly does this mean? Greene's conception of how AB affect determines our intuitive moral judgments closely corresponds to Haidt's (2008) SIM model. As described in Chapter 1, Haidt proposes that our ethical judgments generally stem from the automatic triggering of evolved affective predispositions, based on a perceptual process whereby a novel moral situation is unconsciously matched to an existing moral prototype that carries an emotional charge. We see an event and immediately categorize it, which leads to a powerful emotional reaction that determines the ethical assessments we reach. We subsequently rationalize our intuitive judgments through conscious moral 'reasoning.' This is precisely how

Greene characterizes the production of deontological style assessments via AB affect. He writes, "the emotions hypothesized to drive deontological judgments are…alarm signals that issue simple commands: 'Don't do it!' or 'Must do it!' While such commands can be overridden, they are designed to dominate the decision rather than merely influence it" (pg. 65).

Greene speculates that the influence of SM affect on moral judgment is more subtle. As opposed to AB affect, somatic markers are hypothesized to be constitutive of a conscious moral reasoning process in which different options and outcomes are compared and evaluated through a cost-benefit style analysis. Accordingly, the internal representation of each possible action under consideration carries a unique affective valence, either inclining or disinclining an agent towards picking that option. Some somatic markers will issue greater motivational force than others, privileging the decision-making options to which they are linked. For instance, say that an individual walking down the street observes an armed burglary taking place in a back alley. The witness is safely out of distance and goes unnoticed by both the assailant and victim. The witness quickly considers her options: she can run down the alley and directly confront the assailant, scream "stop" from her current location, call the police from her cell phone, or simply ignore the crime and continue on her way. The consideration of each option will have a unique emotional pull, making it more or less likely to be selected. This appears to be the way Greene conceives of SM affect, emphasizing its role as "currency" in utilitarian, cost-benefit style reasoning. Addressing this point, he writes, "[I] suspect that the consequentialist weighing of harms and benefits is an emotional process…The sorts of emotions hypothesized to be involved here say, 'such and such matters this much.

Factor it in.' In contrast, the emotions hypothesized to drive deontological judgments are far less subtle" (pg. 64).

I have been referring to the type of emotional input that Greene connects with utilitarian judgments as "somatic markers" or SM affect due to the striking similarities between Greene's view and Damasio's "somatic marker theory" of decision-making. According to Damasio, somatic markers operate primarily "as a biasing device" (2000, pg. 174). These markers serve to facilitate decision-making by inhibiting some options, while promoting others. With regards to the former function, somatic markers circumscribe the range of "viable" or "reasonable" decision-making options, as some possibilities are dismissed outright from consideration based on the negative feelings linked to them. Characterizing this role, Damasio writes, "[an inhibitory SM marker] forces attention on the negative outcome to which a given action may lead….The signal may lead you to reject *immediately*, the negative course of action…allowing you to *choose from among fewer alternatives"* (pg. 167). On the other hand, positively charged somatic markers promote the selection of options to which they are linked. "When a positive somatic marker is juxtaposed instead, it becomes a beacon of incentive" (pg. 174). In comparison to Greene, Damasio places more emphasis on the unconscious operation of SM affect and how it sets the stage for—rather than working in conjunction with--conscious moral reasoning. For instance, Damasio writes,

> there is still room for a cost-benefit analysis and proper deductive competence, but only *after* the automated step drastically reduces the number of options. Somatic markers may not be sufficient for normal human decision making since a subsequent process of reasoning and final selection will still take place…[but] somatic markers probably increase the accuracy and efficiency of the decision process (pg. 173).

Elsewhere, Damasio identifies SM affect as the source of moral intuition. He explains, "[somatic markers] may also operate covertly, that is, outside of consciousness…This covert operation would be the source of what we call intuition, the mysterious mechanism by which we arrive at the solution of a problem *without* reasoning toward it" (pg. 187-88).

Both Greene and Damasio endorse psychological noncognitivism (PN)—a thesis that emotion causally influences moral judgment in normal subjects—while agreeing that there is a psychological subset of ethical assessments driven by "alarm bell" affect. Damasio, for example, indicates that there is a unique class of judgments that are linked to emotional markers which function "as an automated alarm signal" (p. 173). On both noncognitivists accounts, the main difference between the class of AB and SM judgments concerns the relative intensity of the emotional input. AB affect is more forceful and, according to Greene's hypothesis, associated with greater activity in "emotional" regions of the brain. In contrast, SM affect issues more subtle urgings that can work in conjunction with a less constrained, conscious moral reasoning process. Recall that Greene strictly classifies AB ethical judgments as deontological in character and SM judgments as utilitarian. He also suggests that AB judgments are primarily intuitive (i.e.,

issued prior to conscious moral reasoning), while SM assessments are more consciously-guided. Figure 3 below displays Greene's taxonomy of the two main psychological types of moral judgment:

Figure 3: Two Psychological Classes of Judgment

Type of Affective Input	Alarm-bell (AB)	Somatic Marker (SM)
Functional Outcome	Deontological	Utilitarian
Processing Style	Intuitive	Reason-based

Damasio seems to have a less restrictive view than Greene. While he never directly addresses the issue, Damasio appears open to the idea that AB affect—which he views as a subtype of SM affect, rather than a unique class—could lead to both utilitarian and deontological judgments. Also, Damasio apparently believes that both AB and SM affect may be linked to either intuitive or reason-based assessments. As I will argue below, I believe that the basic psychological distinction Greene's draws between judgments driven by AB affect and those that are not (SM judgments) is helpful for assessing the plausibility of CPI. I also contend, however, that there are problems with his characterization of these two psychological classes.

For instance, in my opinion, Greene fails to make the case that AB judgments will typically be deontological in character. Again, Greene discovered that contemplation of scenarios, like the Footbridge**,** involving 'up close and personal' harm are correlated with greater activity in 'emotional' brain regions. When asked to judge these types of situations subjects tend to reach deontological conclusions (e.g., it is wrong push the innocent bystander in front of the train). On this basis, Greene concludes that AB

emotion will usually cause subjects to reach deontological conclusions. These findings, however, may be unique to the particular kinds of scenarios utilized in this research--a special class of ethical dilemmas in which there is a relatively clear cut conflict between deontological and utilitarian conclusions. It is plausible that for other types of moral situations, for instance, cases in which saving a life does not require the sacrifice of an innocent person, AB affect could result in utilitarian judgments.

Greene's suggestion that intuitive moral judgments—i.e., assessments that are relatively instantaneous and automatic, and not based on conscious moral reasoning--are typically deontological is also problematic. It seems that, at least under some circumstances, intuitive ethical assessments could yield utilitarian conclusions. For instance, consider a case in which a commercial pilot makes a horrific realization that he or she must make a crash landing and only has a few moments before impact. This pilot then makes a last ditch effort to avoid crashing into a residential area, even though this will almost certainly result in the deaths of everyone on board; whereas there may have been a very minute chance of saving a few passengers by maintaining the plane's current trajectory. This would seem to qualify as an intuitive ethical judgment resulting in a utilitarian conclusion (i.e., choosing the course of action that is likely to save the most lives). Greene indicates that reaching a utilitarian conclusion always requires going through a conscious moral reasoning process, which involves weighing the potential costs and benefits of various options. It is not clear, however, that these kinds of cost-benefit analyses must always be conscious. As argued by Jonah Lehrer (2009), experts in some fields, e.g. veteran pilots, firefighters, etc., appear to reason in this way intuitively.

There is a deeper problem with Greene's attempt to demarcate utilitarian versus deontological judgments merely on the basis of their "functional" difference—i.e., as two alternative kinds of conclusions we might reach in assessing a moral situation. While this functional distinction makes sense for some kinds of ethical cases, like trolley problem scenarios, in which there is a relatively clear conflict between utilitarian and deontological principles, it does not work well for many other types of situations. Indeed, it seems that in many circumstances the same moral conclusion would be consistent with both deontological and utilitarian principles. For instance, the action described above involving the pilot sacrificing herself and her passengers to avoid crashing into a residential area is also compatible with a deontological principle holding that one ought to always avoid harming innocent people when this can be avoided. So is the pilot's decision an example of a utilitarian or a deontological judgment? It seems that it could plausibly be described as both, since in this case utilitarian and deontological principles would likely lead to the same conclusion. This is not a unique case. In many other ethical situations it would not be easy to draw a clear functional distinction between 'the deontological option' and 'the utilitarian option.' Indeed, it is strange that Greene suggests this as a viable approach when he, himself, notes that "most of the standard deontological/Kantian self-characterizations fail to distinguish deontology from other approaches to ethics" (pg. 73). Admitting that it is sometimes difficult to distinguish deontology from utilitarianism, Greene writes that it is a good strategy "to start with concrete disagreements between deontologists and others (such as consequentialists) and then work backward in search of deeper principles" (pg. 74). Unfortunately, Greene fails to identify these deeper principles and instead suggests a functional basis of distinction

that will not work for the many cases in which there is no concrete disagreement between utilitarians and deontologists.

While I am dubious about Greene's claim that intuitive moral judgments cannot result in utilitarian conclusions, another one of his theses seems plausible: in normal subjects, intuitive moral judgment typically involves AB affect. Although it appears unlikely that AB affect only leads to deontological conclusions, it still remains a possibility that this kind of emotional input is constitutive of intuitive moral judgment, whether deontological or utilitarian in character. This view is consistent with Haidt's SIM model of intuitive ethical judgment, which I am broadly endorsing here. Again, according to this model, intuitive ethical assessment typically results from the triggering of an evolved affective predisposition linked to a moral prototype. For example, when we categorize (perhaps unconsciously) a situation as an example of 'cheating,' we naturally have an aversive reaction, and usually judge accordingly. As reviewed in Chapter 1, there is substantial evidence indicating that relatively intense emotion—which we can now identify as AB affect--directly influences and shapes the intuitive moral judgments we reach. For example, various studies by Haidt and colleagues (see Haidt, 2008, for a complete listing) demonstrate that intuitive ethical assessments can be altered (e.g., regarding the perceived seriousness of an ethical violation) by manipulating subjects' disgust responses. Nichols' study (2002b) of disgust norms provides further evidence that AB affect is likely the root cause of our tendency to judge that moral violations are more serious, universal and binding than conventional violations—since we judge violations of disgust norms, which likely involve AB affect, in a similar way

(see next chapter for more details). These studies provide preliminary evidence that, in normal subjects, intuitive ethical judgment will typically involve AB affect.

In offering his "dual processing" theory, Greene also argues that reason-based judgments—i.e., ethical assessments that result from a psychological process in which conscious moral reasoning plays a central role---are generally influenced by SM affect. Based on Damasio's VM patient research (2000), there is strong evidence that, in normal subjects, reason-based ethical judgments will minimally involve SM affect. Again, according to Damasio's "somatic marker" hypothesis, conscious moral reasoning is an emotionally-laden exercise, guided by affective tags or somatic markers that are linked to various decision-making options. We reach ethical decisions based on our background values and commitments, which are manifest in our emotional responses to the representation of various possible outcomes. Accordingly, Damasio hypothesizes that the moral reasoning deficits observed in VM patients are due to a short-circuited somatic marker system. In considering how to act in morally salient situations, members of this clinical population lack the normal emotional responses that typically allow for effective and responsible decision-making. Assuming Damasio's theory is sound, the implication is that in normal populations reason-based ethical judgment will usually involve SM affect, which is consistent with Greene's hypothesis. Recall, however, that Greene makes a stronger claim that reason-based judgment will typically be guided *more* by SM, as opposed to AB affect. This is a more controversial claim, based primarily on Greene's fMRI findings, which as noted above may be specific to the types of moral situations he is testing. It certainly seems plausible to me that a reason-based judgment could ultimately be determined by an AB affective response. Consider cases in which moral

reasoning primarily concerns how to identify a situation (i.e., what moral prototype it exemplifies). Settling on a categorization could trigger an AB affective-response, which might ultimately determine the judgment reached. Hence, while I believe that reason-based judgments typically involve SM affect, I see no reason why these assessments could not involve AB affect as well. Alternatively, it seems plausible that intuitive moral assessments could be influenced by both AB and SM affect.

Despite these concerns, I think it is reasonable, as Greene suggests, to distinguish between ethical judgments that are driven primarily by either AB or SM affect and to loosely identify the former with the psychological class of intuitive ethical assessments and the latter with reason-based assessments. I am more confident about the first point—i.e., there is compelling evidence that some ethical judgments appear to be relatively more 'hot' (AB judgments) than others (SM judgments). For the purpose of evaluating CPI, this basic distinction is vital, since AB judgments likely carry greater motivational force than SM judgments. The contention that AB judgments are typically intuitive in character while SM judgments are characteristically reason-based is more speculative. Nonetheless, this seems plausible as well. Why are intuitive judgments so rapid? Perhaps because this type of judgment is linked to relatively stronger (AB) emotion, which might naturally inhibit conscious moral reasoning, as suggested by Haidt's model. Along similar lines, perhaps engaging in a relatively more 'open,' conscious moral reasoning process, as opposed to *post hoc* rationalization, naturally results in cooler, more SM-based judgments. Again, AB and SM affect are chiefly distinguished by their relative vivacity. Accordingly, it seems reasonable that intuitive judgments would be 'hotter' than reason-based assessments. Clearly, more research must be conducted in this

area before we can confidently endorse this thesis. Nonetheless, it will serve as a working hypothesis for the analysis of CPI that follows in the next section.

Before turning to this question, I want to revisit the psychological cognitivist-noncognitivist debate. As argued in the previous section, the resolution of this issue may have significant implications for the psychological internalist-externalist question. Based on the above analysis, we can consider a revised version of psychological noncognitivism (PN) stipulating that *in normal subjects moral judgment will typically be causally influenced by emotion of either the AB or SM variety, or both*. Both Greene and Damasio appear to endorse this thesis. In contrast, psychological cognitivists, like Marc Hauser (2006), reject PN, proposing instead that at least some moral judgments (e.g., intuitive judgments) are based solely on cold processing. The crux of the debate between psychological noncognitivists and cognitivists concerns at what stage in the moral judgment process emotion enters, since both sides agree that moral judgment and emotion are psychologically connected. Psychological noncognitivists claim that emotion *precedes* and *guides* ethical judgment; while psychological cognitivists propose that, typically, our emotional responses are triggered by antecedent, cold assessments. Thus, the plausibility of PN will depend on whether there is good evidence that either AB or SM affect generally precedes and influences moral judgment in normal subjects. This was just shown above. While there are likely borderline cases that fall somewhere in the middle, most moral judgments can presumably be classified as either intuitive or reason-based. It was argued that either class of assessment is characteristically influenced by a unique type of emotional input, either AB (intuitive judgments) or SM (reason-based judgments)—and there is good reason to suppose that borderline judgments will also

involve one or both of these affective inputs, depending on the psychological systems engaged. Hence, at this early stage in the investigation, the evidence supports PN, as Damasio and Greene suggest. In Section III, I will argue on this basis that CPI likely holds as well. Granting the general assumption that moral motivation is a particular type of desire naturally linked to ethical judgment, it seems that that CPI follows from PN.

To recap, in this section the basic psychological distinction Greene draws between moral judgments that are influenced by AB versus SM affect was endorsed. The main difference between these two kinds of affective input concerns their relative intensity. AB emotion issues stronger impulses that are more likely to lead to the selection or avoidance of a particular decision-making option. While presumably every moral judgment will result from a psychological process involving, at minimum, SM affect, not every assessment will be driven by AB affect. As Greene suggests, it seems likely that there will be substantial overlap between the psychological classes of AB and intuitive ethical judgments on the one hand, and SM and reason-based judgments on the other. Greene also makes some problematic claims, however. He indicates that deontological and utilitarian judgments are distinct psychological classes, each based on a different type of emotional input; and that utilitarian judgments cannot be intuitive in character, since they do not involve AB affect. In response, I argued that it is plausible that utilitarian conclusions may, in fact, be the product of AB affect and can also result from intuitive judgments that do not involve a conscious cost-benefit style of analysis. Moreover, Greene's attempt to functionally differentiate between deontological and utilitarian assessments will not work for the many cases in which these two normative positions would recommend similar courses of action. Finally, I endorsed PN—a thesis stipulating

that *in normal subjects moral judgment will typically be causally influenced by emotion of either the AB or SM variety, or both*—based on evidence that the psychological classes of intuitive and reason-based ethical assessments each characteristically involve one of these two main types of affective input.

III. MAMIT

I argued in the preceding section that empirical evidence supports psychological noncognitivism (PN), a view that, in normal subjects, emotion precedes and influences moral judgment. In reaching this conclusion, I drew two psychological distinctions between AB and SM judgments, on the one hand, and intuitive versus reason-based judgments on the other. It was argued that most ethical assessments can be classified as either intuitive or reason-based, and there is evidence that emotion influences both classes of judgment. Following Greene, I also distinguished between moral judgments that are driven by "alarm-bell" (AB) versus somatic-marker (SM) affect. While there are likely exceptions to the rule, intuitive judgments are characteristically driven by AB affect, while reason-based assessments are typically guided by SM emotion. Since most judgments can be classified as either intuitive or reason-based, PN appears to be true. In this section, I will argue that we can conclude on this basis that internalism likely holds as well.

Within the internalist-externalist literature, moral motivation is generally defined as *an impetus to act in accord with a moral consideration or judgment.* Researchers on both sides of the debate endorse what I am calling the "Moral Affect-Motivation Identity

Theory" (MAMIT), which identifies moral motivation as a particular type of desire. Roskies, for instance, characterizes moral motivation in the following way:

> I take it that [moral] motivation is akin to a species of desire, not necessarily in the sense of intense yearning (my desire for a large portion of French fries), but in the sense sufficient to impel us to action. For instance, although I do not desire to pay my taxes in the same way I desire to eat a large portion of French Fries, I nonetheless am moved to pay them. It is this attenuated form of desire that I intend when I speak of [moral motivation] (2003, pg. 64-65).

According to MAMIT, moral motivation is a particular kind of emotional response that is naturally linked to ethical judgment and reasoning. Based on this view, Roskies argues that the best way to empirically verify the presence of moral motivation is by testing for arousal, utilizing measures such as Skin Conductance Response (SCR) tests. She writes, "I take a measurable SCR to be evidence of the presence of motivation, and lack thereof to be indicative of absence of motivation….The SCR is a reliable indicator of motivation for action" (pg. 57). While questioning Roskies' claims about the reliability of this measure, Kennett and Fine do not take issue with Roskies' underlying commitment to identifying the emotion or arousal linked to ethical judgment and reasoning with moral motivation. Indeed, the central debate in the psychological internalist-externalist literature concerns the truth of psychological noncognitivism—i.e., when precisely emotion enters into the process of moral judgment—and not whether affect is linked with motivation, which is considered a given.

Before evaluating the supposed connection between psychological internalism and noncognitivism, it is worth highlighting that MAMIT accords with the generally

accepted view that weakness of will is a psychologically real phenomenon. As defined in the literature, being morally motivated means that an agent *feels* inclined or disinclined to act on the basis of an ethical consideration or judgment. Accordingly, the intensity or strength of moral motivation can vary in differing contexts. We can feel inclined to act morally, yet succumb to competing desires. As indicated in the quote above, Roskies suggests that our moral motivations will generally suffice for eliciting the action in question, overriding other, nonmoral considerations. Hence, she writes, "failure to act is suggestive, but not proof of, a lack of motivation" (pg. 59). In the next chapter, I will provide an evolutionary explanation for why moral considerations tend to have greater motivational force than many of our nonmoral ones, excluding the nonmoral emotions relating to disgust. For present purposes, the important point is that moral motivation, as defined by MAMIT, is a particular type of desire or emotion, either inclining or disinclining agents towards a particular action. *Contra* the ancient Platonic view, this definition allows for genuine occurrences of weakness of will—i.e., cases where moral motivations fail to elicit the respective moral action. MAMIT makes no commitment, however, with regards to the prevalence and scope of this phenomenon, which is the focus of the contemporary weakness of will debate as expressed in the psychological internalist-externalist literature.

As outlined in Section I, Joyce argues that there is a tight connection between the internalist-externalist and cognitivist-noncognitivist metaethical debates. Indeed, philosophers have generally paired internalism with noncognitivism, and externalism with cognitivism, believing that the truth of one member of the pair implies that of the other. The chief problem with Joyce's proposal is that he fails to distinguish between the

different types of issues (e.g., conceptual, psychological) linked to each general debate. As a result, he does not realize that the psychological noncognitivist-cognitivist issue is the only version of this general debate that is directly relevant to the psychological internalism issue. This brand of internalism proposes that our moral judgments are naturally motivating given our psychological make-up. In normal subjects, judging that action X is morally forbidden (or required) carries with it the motivation to refrain from (or pursue) X. Theorists on both sides of the psychological internalist-externalist issue appear to make the same assumption: if psychological noncognitivism is true, then psychological internalism likely follows. Seemingly everyone agrees that moral emotion carries motivational force. The central debate between psychological internalists and externalists concerns whether emotion precedes and directly influences (psychological noncognitivism) or merely follows (psychological cognitivism) moral judgment. Psychological internalists, like Jesse Prinz (2006), argue that psychological noncognitivism is true and internalism follows because the affective-motivational component of our assessments is part of the judgments themselves. In contrast, psychological externalists, such as Hauser and Roskies, propose that these two systems are separable. Our cold assessments will often, but not always, elicit emotional responses that are motivation-carrying. Hence, based on MAMIT, figuring out when emotion enters into the process of moral judgment is of the utmost importance for determining the truth of psychological internalism.

The presumed connection between psychological internalism and noncognitivism is grounded on MAMIT, a theory that identifies emotion with moral motivation. The basic argument runs as follows: if the moral judgments issued by normal subjects are

typically influenced by emotion, *and emotion carries motivational force*, then these assessments must be motivating. While MAMIT clearly plays a key role in this argument, it would be beyond the scope of this paper to undertake a thorough evaluation of this principle. Rather, following the general trend in the internalist-externalist literature, I will assume that MAMIT holds, with one proviso. I think it is likely that differing types of emotional input will likely carry differing degrees of motivational force. Specifically, based on our analysis, I suspect that AB judgments will typically be more motivating than SM judgments. Again, it is hypothesized that AB emotional input, which is commonly linked to intuitive moral judgments, issues stronger impulses than the SM affect generally connected to reason-based assessments. Accordingly, it seems reasonable that AB judgments would carry relatively greater motivational force. Importantly, although it seems plausible that differing types of emotional input carry varying degrees of moral motivation, this would not undermine MAMIT as a general theory identifying moral motivation with the emotion linked to ethical judgment and reasoning.

My primary goal in this section was to highlight the implications of our finding that psychological noncognitivism is likely true. This implies based on MAMIT—a generally accepted theory identifying emotion with moral motivation--that psychological internalism holds as well. In considering MAMIT, however, I argued that AB judgments likely carry greater motivational force than SM judgments. Accordingly, I will argue in the next section that there is probably less room for *akrasia* in connection with these kinds of judgments, as opposed to cooler assessments based on SM affect alone. Hence,

although it appears that psychological internalism is true, the implications for the weakness of will issue are more nuanced than is commonly supposed.

IV. The Weakness of Will Question

As noted in the opening of this chapter, the psychological internalist-externalist issue is a modern outcropping of the hoary weakness of will debate. In an effort to understand the inner springs of immoral behavior, philosophers have long debated whether such actions are due primarily to poor ethical understanding or *akrasia*. Today, few philosophers deny that weakness of will is a psychologically real phenomenon—i.e., there are occasions when agents know what they morally ought to do, yet fail to act accordingly because other desires overwhelm them. The question remains, however: how widespread are these occurrences? The answer to this question may have important implications for ethical theorizing. One of the traditional goals of moral theorizing is promoting ethical improvement. Historically, at least some moral philosophers have aimed to help people live better lives by first diagnosing the inner sources of immoral behavior and then prescribing methods to overcome them. Accordingly, the type of treatment will vary depending on the diagnosis. For instance, it would be a waste of time emphasizing the importance of strengthening resolve if weakness of will were a very rare occurrence. Rather, there should be more focus on helping people make the right judgments in the first place—enhancing ethical understanding.

While I do not believe that our preliminary findings regarding the truth of psychological internalism can conclusively resolve the contemporary weakness of will debate, these results do provide some insight. Before delving into this, however, I want

to explain why the weakness of will case is not yet closed. As noted in the opening of this chapter, *prima facie*, psychological internalism appears to be more closely aligned with a Platonic view holding that immoral behavior is due primarily to faulty judgment, as opposed to weakness of will. According to this line of thinking, the truth of psychological internalism implies that we are motivated by our judgments, and if the latter holds, then we can assume that people will likely act in accord with their ethical assessments. This, in turn, would suggest that unethical action stems from poor judgment, since people who judge correctly generally follow through and act ethically. The major problem with this argument creeps in with the premise that moral motivation will typically suffice for eliciting the action in question. This assumption relies on a stronger conception of moral motivation than the one endorsed here in connection with the psychological internalist thesis. As defined above, moral motivation refers to *an impetus to act in accord with a moral consideration or judgment.* This would include cases where the inclination is slight or not very intense. Indeed, it was hypothesized above that, as compared to AB emotion, SM affective input carries a weaker motivational force, such that SM judgments leave more room for weakness of will. Hence, given our broad definition, we are not licensed to conclude that moral motivation will likely elicit the action in question. At least in cases where judgments are based solely on SM affect, it seems plausible that an agent could be morally motivated, yet nonetheless succumb to weakness of will. On such occasions, faulty ethical understanding may not be the primary cause of immoral behavior.

The situation is likely different with regards to AB judgments. As noted above, I assume that judgments based on more intense emotional responses carry a relatively

greater motivational force, such that agents will be more likely to act in accord. For this reason, I believe that AB judgments are more motivating than SM judgments. The primary issue with regards to the contemporary weakness of will debate is whether AB judgments generally carry a motivational force *sufficient for eliciting the action in question*. Again, according to the definition here, weakness of will occurs when agents are morally motivated to perform an action but succumb to other desires. It is possible that the motivational force of AB judgments will often suffice for defeating competing desires, such that agents will rarely be led astray after issuing this type of ethical assessment. For this class of judgment, it may be true that morally motivated individuals (i.e., individuals motivated by AB emotion, as opposed to SM) will generally follow through and act accordingly; in which case immoral behavior would be due to faulty judgment, rather than *akrasia*, since issuing a morally correct judgment would naturally result in ethical action. Hence, the implications would be significant if it were, in fact, the case that ethical assessments driven by AB affect typically suffice for eliciting the action in question. While this is technically an empirical question, developing reliable methods for testing this hypothesis would be a great challenge. At this point, the best we have is indirect evidence.

As noted above, I assume that the class of AB judgments overlaps substantially with the class of intuitive (as opposed to reason-based) ethical assessments. In general, AB judgments are rapid, automatic and do not rely on conscious moral reasoning. As outlined in Chapter 1, the findings from Haidt and colleagues (for a review, see Haidt & Bjorklund, 2008) regarding the prevalence of "moral dumbfounding" indicate that we generally do not reflect upon the legitimacy of our intuitive judgments, and even when

forced to question them, we still do not relinquish these views. It appears that we generally remain committed to the veracity of our strong gut feelings—which would seem to increase the likelihood that we will act upon judgments that are driven by them. Based on this evidence and the reasonable hypothesis that stronger emotional responses carry a relatively greater motivational force, there seems to be little doubt that intuitive ethical judgments are typically more motivating than reason-based assessments. The former class is usually driven by intense AB affect, while the latter class is characteristically influenced by relatively more subtle SM emotion. Nonetheless, it would be premature to conclude that intuitive assessments generally leave no room for weakness of will. It seems plausible that agents could make 'snap judgments' on the basis of AB affect, yet nonetheless fail to act accordingly. The prevalence of such occurrences remains an open question, but presumably the proportion of *akrasia* cases is much lower for intuitive, AB judgments as compared to reason-based, SM judgments. Generally speaking, it appears that the latter class of judgments leaves more room for weakness of will.

Hence, our finding that psychological internalism is true offers some new insight regarding the contemporary weakness of will debate. This is a more complicated matter than some theorists have assumed. There are likely different types and degrees of ethical motivation—e.g., motivation based on alarm-bell versus somatic-marker affect—and we should not suppose that the truth of psychological internalism automatically rules out the possibility that *akrasia* is a common occurrence. In order to be morally motivated, agents need only to feel inclined to act in accord with an ethical consideration or assessment. This still leaves room for cases in which agents are morally motivated yet succumb to

competing desires that are stronger. Indeed, the central issue in connection with the contemporary weakness of will debate is not whether agents are motivated by their judgments or not—apparently they are, since psychological internalism holds--but rather if these motivations will typically suffice for eliciting the action in question. Only this stronger thesis would imply that immoral behavior generally stems from poor ethical understanding, rather than weakness of will. There is currently insufficient evidence to endorse this strong thesis. In the next chapter, I will argue on evolutionary and psychological grounds that moral considerations are typically afforded greater weight than nonmoral ones in practical decision-making—but this still falls short of establishing that moral motivations will generally elicit congruent behavior. Even though we naturally afford moral norms greater practical clout, we may nonetheless fail to choose the 'ethical course of action' based on a variety of considerations (e.g., situational pressures, personality variables, competing motivations, etc.).

The above analysis also points to a major limitation in the conventional approach to the weakness of will question. In writing about this issue, theorists typically fail to distinguish between differing psychological classes of judgment. Instead, there is an unfortunate tendency to group disparate types of moral judgments together—a mistake closely paralleling the one that has hindered empirical progress on the psychological internalism-externalism question: the failure to separate distinct populations of subjects for the purpose of evaluating metaethical questions regarding the nature of moral judgment. With regard to the contemporary weakness of will debate, it may very well be the case that differing types of moral judgment leave relatively more or less room for weakness of will. Since AB judgments are based on stronger emotional impulses, it

seems likely that they will generally carry greater motivational force than SM assessments. Hence, agents should find it harder to resist the behavioral impulses linked to AB judgments. As noted above, there is substantial overlap between the psychological classes of AB and intuitive moral judgments, on the one hand, and SM and reason-based judgments on the other. Based on the investigation here, I speculate that for normal subjects akrasia occurs substantially less frequently in the case of intuitive moral judgments, as opposed to reason-based assessments, which are based more on conscious moral reasoning and weaker SM affect. This would imply that faulty ethical understanding or misperception, as opposed to weakness of will, is likely a more common source of immoral behavior in connection with intuitive judgment. At least for this class of assessment, our initial response may be the key factor that determines whether we act ethically or not, since we seem to generally follow our strong gut feelings.

Of course, given the limited range of evidence, these conclusions are speculative. Nonetheless, while the precise differences remain somewhat mysterious, there seems to be little doubt that unique psychological classes of moral judgment carry varying degrees of motivational force, such that the prevalence of *akratic* occurrences will differ from class to class. Ethicists focused on enhancing moral behavior in the general population may need to tailor their message accordingly. For instance, it seems that recommendations regarding how to strengthen resolve and resist temptation would more effectively target immoral behavior that stems from reason-based moral judgments. On the other hand, lessons about how to improve moral understanding and perception may be the best means of enhancing behavior that follows from intuitive judgments. Again,

given the current evidence, these are merely provisional theses. They reflect, however, the more nuanced implications and questions that can follow from this new kind of approach to the weakness of will issue—a style of analysis that carefully distinguishes between unique psychological classes of judgment as well as distinct populations of subjects.

In this chapter, a similar method was applied to the psychological internalist-externalist issue, with a good measure of success. Bucking a misguided trend in the literature, I limited my focus to "normally functioning" subjects, broadly construed. In contrast, as exemplified by Roskies' writing, the standard practice involves identifying a special population (e.g., psychopaths or VM patients) that supposedly serves as a "walking counterexample" to the internalist thesis; and then arguing on this basis that internalism *writ large* is false. It appears, however, that these special populations are not, in fact, counterexamples to the internalist thesis, since they appear to lack a capacity for genuine moral judgment, as measured by Kennett and Fine's test. Regardless, this general, all-or-nothing approach to evaluating descriptive metaethical theses seems wrong-headed in itself: the expression of an outmoded tendency to view such questions as metaphysical, rather than psychological, in nature. Even if a walking counterexample to the internalist thesis could be found, would this provide good reason to reject psychological internalism *tout court*? The answer to this question depends on the goal of descriptive metaethical theorizing. If our aim is to discover universal and necessary metaethical truths--assuming it makes sense to speak in such terms, which is not at all clear to me—then we should endorse the conventional approach. On the other hand, if we want to learn more about the complexity of human moral psychology and the inner

springs of ethical action, this would call for a more refined method of the sort utilized here. Seeking to uncover the unique characteristics of distinctive populations is undoubtedly a valuable scientific endeavor, but hastily generalizing these findings may carry a great cost. By avoiding this pitfall, we discovered that the internalist thesis likely applies to normal subjects, albeit in a nuanced way. This metaethical pronouncement certainly lacks the grandiosity of traditional metaphysical claims--which is probably a good thing, given that the conventional approach has generally led only to stalemates.

Chapter 3 References

Augustine. (1961). *Confessions.* Trans. R.S. Pine-coffin. Penguin Classics.

Blair, R.J.R., & Cipolotti, L. (2000). Impaired social response reversal: A case of "acquired sociopathy." *Brain*, 123: 1122-1141.

Damasio, A. R. (2000). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Random House.

Greene, J. (2008). The secret joke of Kant's soul. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: Harper Collins.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814-834.

Haidt, J., & Bjorklund F. (2008). Social Intuitionists Answer Six Questions about Moral Psychology. In *Moral Psychology, Volume 2* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Hare, R. D. *Without conscience: The disturbing world of the psychopaths among us.* New York: Pocket Books.

Joyce, R. (2008). What Neuroscience Can (and Cannot) Contribute to Metaethics. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Kant, I. [1785] (1964). *Groundwork of the Metaphysics of Morals*, Trans. H.J. Patton. New York: Harper Torchbooks.

Kennett, J., & Fine, C. (2008). Internalism and the Evidence from Pscyhopaths and "Acquired Sociopaths." In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Nichols, S. (2002a). Is it irrational to be immoral. How psychopaths threaten moral rationalism. *Cognition,* 84: 285-304.

Nichols, S. (2002b). Norms with feeling towards a psychological account of moral judgment. *Cognition* 84: 221-236.

Nichols, S. (2004a). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford UP.

Nichols, S. (2008). Moral Rationalism and Empirical Immunity. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Plato. (1981). *Five Dialogues.* Trans. G.M.A. Grube. Indianapolis: Hackett.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations,* 9: 29-43.

Roskies, A. (2003). Are ethical judgments intrinsically motivational? Lessons from "acquired sociopathy." *Philosophical Psychology,* 16: 51-66.

Roskies, A. (2008). Internalism and the Evidence from Pathology. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Smith, M. (2008). The Truth about Internalism. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Young, L. Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neurophysical evidence. *Journal of Cognition and Culture*, 6: 265-278.

4. The Affective Foundations of Practical Clout: A Naturalistic Critique of Moral Error Theory

"Vice and virtue…are not qualities in objects, but perceptions in the mind: And this discovery in morals…is to be regarded as a considerable advancement of the speculative sciences…tho it has little or no influence on practice. Nothing can be more real or concern us more than our sentiments…[and] no more can be requisite for the regulation of our conduct and behavior."

-David Hume, *Treatise*

In the late 1970s, John Mackie popularized and expounded "moral error theory," a sceptical position inspired by Hume. In his *Treatise*, Hume (1967) famously underscores the human propensity to project our feelings and perceptions upon the world, falsely attributing these qualities to external objects, rather than recognizing their true source—our distinctive psychology. This tendency, Hume theorizes, extends to the moral domain, where we unreflectively suppose that ethical behaviors and situations are objectively right or wrong, independent of our subjective feelings and preferences. He insists that this natural belief is mistaken. Extending Hume's basic argument, Mackie concludes that our everyday moral judgments are "false," since they "include a claim to objectivity"—which philosophical reflection reveals to be erroneous. Mackie writes, "moral scepticism must…take the form of an error theory, admitting that a belief in objective values is built into ordinary moral thought and language, but holding that this ingrained belief is false" (1977, p. 48-9).

Evolutionary moral error theory (EMET) is a recent offshoot of Mackie's original theory. Theorists, such as Michael Ruse (1986) and Richard Joyce (2006), have worked to supplement Mackie's basic position with an evolutionary account of our 'collective moral illusion.' Mackie's original theory leaves an important question unanswered: why is this tendency to 'objectify' morality part of our psychological make-up? According to

EMET, we have this natural tendency because it was adaptive for our ancestors. By objectifying our judgments in this way, there is a greater likelihood that we will act upon them. Ruse writes, "the evolutionist points out why it is part of our nature to objectify morality. If we did not regard it as binding, we would ignore it. It is precisely because we think that morality is more than mere subjective desires, that we are led to obey it" (p. 103).

Like Mackie's original error theory, EMET rests on a dubious psychological assumption. Error theorists presuppose that we are naturally, psychologically committed to 'moral objectivism' or 'realism[2],' in the more traditional sense of the terms (see Section I for a discussion of the slippery nature of this philosophical concept)—i.e., a thesis that *there are 'independent' moral truths or facts that apply irrespective of our subjective preferences.* According to moral error theory, since we have this natural belief, our moral judgments include a (often tacit) claim to objective truth. Consider the following two propositional forms:

Relativist moral judgment*; 'action X is morally wrong.'*

Objectivist moral judgment; *'objectively, action X is morally wrong.'*

Error theorists contend that our judgments have the latter form, and *in this sense*, they are erroneous, since moral objectivism is false. Error enters in with the qualification that, *objectively*, action X is ethically wrong, as opposed to merely claiming that it is ethically wrong in a relativist sense of the term.[3] Hence, error theory is not as radical as it may sound. In denying that there are objective moral truths or facts, error theorists are *not*

---

[2] In this chapter, I use the terms 'moral objectivism" and "moral realism" interchangeably.

3For our purposes, I define 'ethical relativism' negatively, as a general metaethical position *denying* that moral objectivism or realism holds. Ethical relativists reject the claim that there are 'independent' moral truths or fact that apply irrespective of our subjective preferences. Just as there are many brands of objectivism (see discussion below), relativism encompasses a range of views as well.

thereby committed to a view that our moral judgments are normatively unjustifiable—only that they cannot be justified by an appeal to objectivist truths or facts. The metaethical question concerning whether or not ethical relativism is a normatively adequate position, which has been the subject of much debate, will not be pursued here. Instead, I will focus primarily on the empirical question of whether we are, as the error theorists claim, naturally committed to moral objectivism.

Indeed, I will argue that this psychological thesis lacks empirical support, a finding that undermines moral error theory by challenging its central claim that everyday moral judgment includes an (erroneous) objectivist appeal. With regard to relevant empirical findings, researchers have mistakenly supposed that psychological studies focusing on how we distinguish between moral and conventional normative violations establish that we are inherently ethical objectivists. I contend that these studies support only a weaker thesis that we naturally invest morality with what Joyce refers to as "practical clout." That is, we tend to attribute greater weight to moral reasons in practical deliberation, such that they will often trump nonmoral considerations. The evidence suggests that this proclivity is grounded in affect, as opposed to a belief in moral realism. This finding, in turn, should allay traditional fears that the adoption of a relativist belief system will typically lead to an upsurge of immoral behavior.

I. The Case Against Moral Objectivism

Moral error theories, including EMET, rely on two main premises, which are not mutually entailing. One thesis is psychological in nature—i.e., the claim that we are inherently moral objectivists. As I will argue in the next section, this thesis lacks

empirical support. The other primary thesis of moral error theory is that endorsing moral objectivism is epistemologically unjustified. Combining these two theses, one psychological and the other metaethical, moral error theorists argue that our everyday moral judgments are false *since they include an unjustified claim to objectivist truth.* In this section, I evaluate the metaethical claim that a belief in moral objectivism is without epistemological warrant, since this plays a key role in the basic argument for moral error theory. Indeed, I will contend that, with regard to this thesis, moral error theory is on firmer ground. Again, just to be clear, the question at issue here is whether or not a *belief* in moral realism can be *justified*. Are there good reasons for believing that this thesis is true? It is possible that a proposition could be true, despite the fact that we lack good reasons for believing this. Hence, in arguing that a Darwinian genealogy of our moral sensibility undermines (a belief in) moral realism[4], I am not thereby committed to the stronger claim that moral realism is false. My claim is only that an endorsement of moral objectivism is *not epistemologically justified*.

As defined here, moral 'objectivism' or 'realism' is a metaethical thesis stipulating *that there are 'independent' ethical truths or facts that apply irrespective of our subjective preferences.* I have been referring to this thesis as "the traditional view" since there is a great variety of 'realist' moral theories, and some of the more contemporary, naturalistic versions do not qualify under the definition endorsed here. Similarly, due to a variation in views that do qualify, in our broad definition of moral objectivism, I have set off the term 'independent' in reference to the moral truths or facts

[4] For ease of reading, I will use the phrases "undermines moral realism" and "argues against moral objectivism" as short-forms for "undermines *a belief* in moral realism" and "argues against *a belief* in moral realism," respectively, which are the meanings I intend.

postulated. Indeed, on our account, there are two main types of realist views, Platonic and Kantian, each positing a different type of independence relation. Plato famously argued that there are *independently-existing*, divine Forms, which are the source of ethical truths and values. Accordingly, Platonic versions of moral objectivism are more metaphysically-oriented. Divine Command theory (i.e., a metaethical position stipulating that God's will determines what is moral or immoral), which is still espoused in contemporary religious settings, is a prominent example of this type of objectivist view. In comparison, Kantian realist views, which enjoy greater popularity among contemporary ethicists, are more epistemologically-based. Kant contended that moral rules are ultimately based on "Universal Reason," rational rules which apply irrespective of our subjective preferences and are irreducible to scientific laws and processes. Correspondingly, as generally conceived here, Kantian *realist* theories[5] postulate that there are ethical truths, facts or rules that are not subject to a naturalistic explanation. Contemporary examples of Kantian realist views are provided by Thomas Nagel (1986) and Russ Shafer-Landau (2003).

While our definition of moral realism or objectivism—which stipulates, *there are 'independent' ethical truths or facts that apply irrespective of our subjective preferences*--encompasses both Platonic and Kantian versions, it excludes another kind of 'realist' account. In recent years, a small number of moral philosophers, such as Richard Boyd (1997) and Peter Railton (1997), have proposed that natural facts or properties (e.g., facts about how to fulfill our basic desires) are constitutive of 'objective' moral truths. These

[5] Of note: I am not claiming that all contemporary 'Kantian' views, broadly construed, are realist in nature. Only Kantian moral theories postulating that there are non-natural moral facts, values or rules qualify under this account. Accordingly, the Kantian "constructivist" views of John Rawls and Christine Korsgaard are not realist views on this account.

'moral naturalist' views fall outside of the scope of moral realism, as defined here, since they assert that the moral truths they postulate *depend* on contingent facts about human nature and our subjective preferences—in a way that Kantian and Platonic realists would find unacceptable. Clearly, there are a great variety of moral 'realist' positions, each positing a different type of independence relation. Three main types have been identified here, Platonic, Kantian and Naturalistic. As classified here (see Street, 2006, for a similar taxonomy), moral realism or objectivism includes only the more traditional Platonic and Kantian notions. These two main types of realist views are the chief targets of moral error theorists, such as Ruse and Joyce, which is why I am utilizing this particular definition of moral objectivism. Accordingly, when moral error theorists make the empirical claim that we are naturally committed to moral objectivism, it is in this general sense of the term. As I will show in the next section, there is scant empirical evidence in support of this psychological thesis.

The other main claim of moral error theorists—i.e., that endorsing ethical objectivism is epistemologically unjustified—fares better. There are good arguments in support of this claim. As both Ruse and Joyce contend, an evolutionary genealogy of our moral sensibility, of the sort endorsed in this project, threatens to undermine moral realism. In short, within this broad evolutionary framework, moral objectivism appears superfluous. It appears that we can fully explain the recurrent patterns in moral cognition—as well as our tendency to imbue moral norms with practical clout--without reference to objective moral truths. Many of the intuitions at the core of our ethical judgments and reasoning are likely the product of natural selection. These biases and predispositions were adaptive for our ancestors, promoting their genetic fitness, and this

is why they are part of our cognitive make-up today. Accordingly, from an evolutionary perspective, our tendency to impute practical clout to moral norms—i.e., attribute greater weight to moral considerations in practical deliberation, such that they will often trump nonmoral considerations—is an adaptive 'binding' device, increasing the likelihood that we will act in accord with our evolved ethical intuitions. Summarizing this Darwinian challenge to moral realism, Joyce writes, "we have an empirically confirmed theory about where our moral judgments come from (we are supposing). This [theory] doesn't state or imply that they are [objectively] true…They *could* be true, but we have no reason for thinking so. (p. 211).

The argument against accepting Platonic versions of moral realism involves a relatively straightforward appeal to Ockham's Razor. As noted above, Platonic versions of moral realism postulate the independent *existence* of moral truths or values, which are not subject to natural laws or facts. Obviously, from a scientific perspective, this type of conception poses many difficulties. Addressing this issue in his famous "argument from queerness," Mackie writes,

> if there were objective values [of the sort postulated by Platonic realists], then they would be entities or qualities of a very strange sort, utterly different from anything else in the universe…How much simpler and more comprehensible the situation would be if we could replace the [objective] moral quality with some sort of subjective response (p. 38- 40).

Ruse offers a similar argument against endorsing Platonic realism. "The evolutionist's claim is…that morality is subjective—it is all a question of human feelings and sentiments….In the light of what we know of evolutionary processes, the objective

foundation has to be judged *redundant*" (p. 102-108, emphasis mine). Ruse contends that this Darwinian genealogy provides a complete, scientific explanation of our fundamental moral biases and beliefs, and introducing a non-naturalistic ontology (i.e., independently-existing moral values) into this picture introduces needless difficulties. As compared to this Darwinian theory, Platonic realism is less clear and parsimonious and lacking in explanatory power. Hence, we ought to endorse the former and reject the latter.

It would be beyond our scope to offer a detailed evaluation of this argument against endorsing Platonic realism. From a scientific perspective, it appears very strong. Of course, contemporary proponents of Platonic realism typically deny that this naturalistic perspective is the only epistemologically valid standpoint to adopt, contending that religious faith or revelation can be another source of knowledge. I do not agree with this contention, but I will not argue the point here. Indeed, a more interesting question is whether a similar argument to the one Ruse deploys against Platonic objectivism also works against Kantian realism, which is a more popular view in contemporary moral philosophy. Indeed, one reason that modern moral philosophers typically prefer this more epistemologically-oriented realist view is that it avoids positing the extra ontology characteristic of Platonic versions, which are so difficult to scientifically explain. Kantian moral realists contend that the 'independent' moral truths they postulate are not ontological entities, but rather, rational 'laws' or rules—and hence, this type of realist view is immune to Ockham's Razor.

Sharon Street (1996), however, forcefully argues that a Darwinian genealogy of our moral sensibility also undermines Kantian versions of moral realism. Street argues

that moral realists, either Platonic or Kantian, who accept that our fundamental ethical attitudes are the product of natural selection face a challenge: they must explain how the Darwinian process that shaped our moral sensibility relates to the objective moral truths they postulate. Street claims that moral realists have only two options—they can either deny that there was any relationship or claim that our evolved attitudes were adaptive because they are truth-tracking—both of which are unsatisfactory. The 'no relation' option leads to the implausible conclusion that our moral attitudes and beliefs are likely false, not just with regard to their objectivist trappings, as the moral error theorists claim, but in their very content (e.g., it is not true that murder is 'wrong,' in any sense of the term). The tracking account, on the other hand, fails to meet the standards of good scientific explanation. Hence, Street contends that moral realism is struck down by either horn of an inescapable "Darwinian dilemma."

She underscores that this dilemma for the moral realist follows from the evolutionary genealogy of our basic moral attitudes and beliefs. As noted above, the primary implication of this account is that our moral sensibility is designed to promote reproductive success, which does not entail that it is truth-tracking. Assuming that they accept this genealogy, moral realists must take a stand regarding this issue: was the Darwinian process that shaped our evolved ethical attitudes directly related to these objective moral truths, or not? Was it perhaps the case that these proclivities were adaptive *because they were truth-tracking*? According to Street, if realists answer 'no' to these questions, insisting that this selection process unfolded irrespective of the values they postulate, they are left in a very difficult spot. "The key point to see about this option is that if one takes it, then the forces of natural selection must be viewed as a

purely distorting influence on our evaluative judgments, having pushed us in evaluative directions that have nothing whatsoever to do with evaluative truth" (p. 121). Street grants that while it is a logical possibility that these blind forces could have, purely by chance, led to a convergence between our evolved intuitions and objective truth, this seems overwhelmingly unlikely given all the possibilities for error along the way. Hence, Street argues that this 'no relation' option leads to "the implausible conclusion that our evaluative judgments are in all likelihood mostly off track" (p. 122).

Street acknowledges that a Kantian moral realist may attempt to parry this horn of the dilemma by citing the power of rational reflection to overcome these evolved biases. A Kantian might insist that Reason has the power to stand above these Darwinian influences and provide an objective means of evaluating and correcting our unreflective biases. If this strategy succeeds, it would explain how we can nonetheless discover objective moral truths even though the process of natural selection bore no direct relation to them. Street responds to this objection by insisting that it relies on an unrealistic model of rational reflection, one falsely implying the possibility of standing apart from any evaluative standpoint, a view from nowhere, as it were. She writes, "but this [rationalist] picture cannot be right. For what rational reflection about evaluative matters involves, inescapably, is assessing some evaluative judgments in terms of others" (p. 124). Street contends that, since our evolved judgments can only be rationally evaluated from a Darwinian evaluative standpoint (i.e., by relying on other evolved intutions), there is simply no escaping the sceptical implications of the 'no relation' position. She emphasizes, "if the fund of evaluative judgments was thoroughly contaminated with

illegitimate influence—and the objector has offered no reason to doubt *this* part of the argument—then the tools of rational reflection were equally contaminated" (p. 124).

According to Street, the only other option available to the moral realist who endorses an evolutionary genealogy of our core moral intuitions is to posit a direct causal relationship between Darwinian selection pressures and the objective truths of ethics. Specifically, the realist may claim that our evolved intuitions were adaptive *because they were truth-tracking*. On this view, having an ability to discern moral facts conferred a reproductive advantage, and this explains why our ethical proclivities accurately reflect moral truth. Characterizing this alternative, Street writes, "the evaluative judgments that it proved most selectively advantageous to make are, in general, precisely those evaluative judgments which are true" (p. 125). Street argues that, based on the standards of good scientific explanation, this "tracking account" of our evolved moral sensibilities is clearly inferior to the alternative, "adaptive link" theory. On the latter account, which makes no reference to the discovery of objective truths, our moral biases were adaptive simply because they motivated behaviors conducive to reproductive success—"forged adaptive links between our ancestors' circumstances and their responses to those circumstances" (p. 127).

Street contends that this adaptive link theory is more parsimonious, clear and powerful than the tracking option. The parsimony point is straightforward. The tracking account postulates objective moral truths, whereas the adaptive link account does not. With regard to the clarity and explanatory power issue, Street argues that the tracking account also falls short. Specifically, it fails to proffer an adequate explanation of why perceiving moral truths—as opposed to merely believing (mistakenly) that our

judgments are objectively true--would be fitness enhancing. Unlike other types of beliefs or judgments where a link between accurate perception and reproductive success seems more vital—e.g., beliefs pertaining to basic causal relations in nature or simple mathematical operations—this connection is not at all obvious in the case of moral beliefs. Indeed, the adaptive link account explains why our judgments are fitness enhancing, without any reference to moral truth. Street writes, "[how can a realist] tracking account explain the remarkable coincidence that so many of the truths it posits turn out to be exactly the same judgments that forge adaptive links…---the very same judgments we would expect to see if our judgments had been selected on those grounds alone, regardless of their truth" (p. 132)? This challenge for the moral realist grows even greater in considering the haphazard nature of the environmental influences that shaped our moral biases: had the environment been substantially different, a different range of ethical attitudes would have been adaptive. Street insists that a realist tracking account[6] cannot answer these challengess, giving us good reason to endorse the alternative, adaptive link theory, which provides a more parsimonious and illuminating explanation of our 'ethical' proclivities.

[6] Interestingly, Street argues that 'moral naturalist' versions of realism, which have been excluded from consideration here, are also vulnerable to this critique. The general thesis of moral naturalism is that the evaluative truths or facts of morality are identical to or constituted by natural facts about human beings. It is plausible that correctly identifying such facts could promote survival. Hence, for this type of realist view, a link between truth-tracking perception and reproductive success may be easier to establish. Street answers this objection by emphasizing that such a position only reintroduces the same basic dilemma at another level. The moral naturalist must now explain how we can reliably identify these fact-value identities. How was this capacity related to Darwinian selection pressures? The 'no relation' option leads to the same problems cited above (i.e. there is no reflective standpoint independent of Darwinian influence). With regards to the tracking option, Street writes, "it is even more obscure [than the initial tracking hypothesis] how tracking something as esoteric as independent facts about natural-normative identities could ever have promoted reproductive success" (p. 141). While the naturalist could plausibly argue for the evolutionary benefit of recognizing some truths about human nature, there is no comparable case to be made for an ability to correctly identify these truths as normative facts. Thus, Street insists that any realist tracking account confronts the same basic difficulty.

Street's argument relies on the assumption that no moral realist would seriously propose that most of our core moral beliefs are false. If this premise is denied, then her position loses much of its force. Street rejects the 'no relation' option for moral realism because it would imply the "implausible" conclusion that our intuitive judgments are mostly off-track, since it is highly unlikely that a blind evolutionary process bearing no direct relation to the objective truths of morality would result in truth-tracking intuitions. Of course, this conclusion is only implausible if one is committed to the view that our evolved attitudes cannot be mostly false. Street's answer to the rationalist objection to her argument also relies on this presupposition. As noted above, Kantian rationalists might contend that, even though Darwinian influence 'contaminated' our unreflective attitudes, reason affords us the power to impartially evaluate and correct these evolved biases. Street responds that this kind of moral reflection must take place from some evaluative standpoint (i.e., be anchored by a set of core intuitions), which will, itself, be 'contaminated' by Darwinian influence. The plausibility of this claim would seem to depend on the content of the normative theory being proposed. If there appears to be substantial overlap between the objective moral truths prescribed and our Darwinian intuitions, the claim that the former were discovered or generated without reliance on the latter would seem highly dubious. On the other hand, if a realist normative theory suggested a radically counter-intuitive view of moral obligation, it could very well be the case that this rationalist view was untainted by Darwinian influence. Street's point is that we just do not see these types of normative views being seriously proposed. Hence, her argument against moral realism is hypothetical: if you are a moral realist committed

to a Darwinian genealogy of our core moral beliefs *and you are not proposing a radically counter-intuitive normative theory*, then you face the following dilemma…

How should we evaluate Street's assumption that no Kantian moral realist (or any other type of realist, for that matter) would earnestly recommend that most of our moral intuitions are mistaken? Street's contention seems to be based on historical observation. Her argument runs as follows: surveying the Kantian moral realist theories of the past, it appears that the vast majority have recommended normative theories that align, at least in large part, with our intuitive beliefs. Hence, it seems that Kantian moral realists are generally committed to the soundness of (most of) our ethical intuitions. In his critique of deontological philosophy, Joshua Greene (2008) makes a similar observation, writing "very few philosophers are actively challenging anyone's moral intuitions" (p. 75). While I cannot offer a detailed analysis here, I believe that Street and Greene's observation, generally speaking, is sound. Although there may be elements of a Kantian realist theory that are counter-intuitive (e.g., Kant's claim that we should never tell a lie, under any circumstances), I cannot think of any theory that is, on the whole, radically at odds with most of our evolved intuitions. What would a radically counter-intuitive moral theory look like? Street offers a nice example (pg. 116). Consider the following theory, which stipulates:

- The fact that something would promote one's survival is a reason against it.
- The fact that something would promote the interests of a family member is not a reason to do it.
- The fact that someone has treated one well is a reason to do that individual harm in return.
- The fact that someone is altruistic is a reason to dislike, condemn, and punish him or her.
- The fact that someone has done deliberate harm is a reason to seek out that person's company and reward him or her.

This moral theory would clearly run afoul of our evolved intuitions. The fact that it seems so absurd reinforces Street's basic contention: we simply do not find Kantian moral realists proposing radically counter-intuitive theories, like the example above. Instead, Kantian moral realists, and other types of realists as well, typically offer normative theories that largely align with *most* of our basic intuitions.

Hence, as the proponents of EMET contend, a Darwinian account of our moral sensibility poses a significant epistemological challenge to moral realism, both Platonic and Kantian varieties. Like Ruse argues, Platonic realism, which posits a scientifically implausible account of independently-existing sources of moral value, falls victim to Ockham's razor: the alternative, Darwinian explanation of our core moral beliefs and tendency to impute practical clout to moral norms is more parsimonious, clear and has greater explanatory power. Similar epistemological considerations weigh against an endorsement of Kantian versions of moral realism. If Kantian moral realists accept a Darwinian genealogy of our core moral attitudes (which they should) while proposing a normative theory that largely accords with our evolved intuitions (which they all seem to do), they face a real problem: as a scientific explanation of our evolved ethical attitudes, realist tracking accounts are much less compelling than the alternative, adaptive link theory. Hence, it appears that Ruse and Joyce are correct: an evolutionary explanation of our core moral intuitions undermines moral realism. Given the countervailing, Darwinian arguments, it appears that an endorsement of ethical objectivism is not epistemologically justified.

II. Are We Really Moral Objectivists?

In the preceding section, it was argued that an evolutionary genealogy of our moral sensibility undermines moral realism or objectivism—a metaethical position postulating that there are 'independent' moral truths or facts that apply irrespective of our subjective preferences. In line with this contention, evolutionary moral error theorists, such as Ruse and Mackie, propose that our everyday moral judgments are false, since these judgments include an unjustified claim to objectivist truth. This argument, however, relies on a dubious psychological assumption: that we are inherently ethical objectivists, in the sense defined above. As it turns out, there is scant experimental evidence that moral objectivism is our default setting. As Nichols (2004b) underscores, studies of how we distinguish between moral and conventional normative violations--despite what many of the researchers involved in this work suppose--fail to establish this thesis. The strongest experimental evidence in support of the claim that we are naturally moral objectivists comes from a small set of studies suggesting that young children may have realist leanings, but there are holes in this research as well. In contrast, studies (Nichols, 2004b) reveal that substantial proportions of college students are apparently ethical relativists. Taken together, the available evidence provides insufficient support for the moral error theorists' claim that ethical objectivism is our default setting. As such, it appears that moral error theory rests on an unwarranted psychological assumption.

There have been numerous studies focusing on how we naturally distinguish between moral and conventional normative violations (see Smetana, 1993, for review).

The four most common measures utilized in this research are: perceived seriousness of the violation, the permissibility of violating the norm, whether or not the norm's legitimacy is dependent on social authority and the type of justification offered for the norm. These questionnaire or interview-based studies sample subjects' responses to target scenarios. Moral scenarios typically describe cases involving pain and suffering (e.g., one child hits another without provocation), while conventional scenarios focus more on etiquette violations (e.g., talking out of turn in the classroom). Across studies, results show that normal subjects tend to view moral normative violations as less permissible and authority-contingent and more serious than conventional violations. Moreover, moral norms are generally given welfare-based justifications, as opposed to conventional norms, which typically receive social/conventional explanations. The general capacity to distinguish between normative and conventional violations along these core dimensions has been found in children as young as three and a half years old (Turiel, 1983). As reported by Blair (1995), psychopaths, in contrast, lack this ability, as they tend to treat all violations as conventional (see Chapter 2 for more details).

Researchers involved in these studies have generally interpreted their findings as indicating that, unlike in the case of conventional norms, subjects view moral norms in *objectivist terms*.[7] These researchers infer that, since normal subjects tend to judge moral norms as being less permissible and authority contingent and more serious than conventional violations, these subjects are ethical objectivists. Nichols (2002a) forcefully challenges this interpretation of the data, emphasizing that subjects could offer nonconventional responses without a concomitant belief in moral realism. Believing that

[7] Nucci (2000), for example, concludes, "pre-school aged children understand that it is objectively wrong to hurt others" (p. 86); sourced from Nichols (2004a).

moral norms are more serious and generalizable, and not authority-contingent is probably necessary—but certainly not sufficient—to qualify as a moral objectivist; which would require an additional commitment to these norms applying irrespective of our subjective preferences. Addressing this point, Nichols writes, "the prevailing measure for moral judgment, the moral/conventional task, does a poor job of assessing whether [subjects] regard moral properties as dependent on our responses" (2002a, p. 173).

Indeed, in five studies with college student participants, Nichols (2004b) discovered that 'ethical relativists' draw the typical distinction between moral and conventional normative violations. In these experiments, subjects were given a revised version of the standard moral/conventional test. Before they were instructed to judge the target scenarios along the four typical dimensions (i.e., permissibility, authority contingency, seriousness and justification type), subjects were presented with conflicting opinions regarding the normative status of a given scenario (e.g., Frank says, "it is ok to hit others when you feel like it;" Bob says, "it is wrong to hit others when you feel like it") and were asked whether there was a "fact of the matter" to settle to this dispute. Subjects who consistently denied that there was a fact of the matter to settle moral disputes were classified as "ethical relativists." The proportion of subjects fitting this classification varied across the five studies, from a low of approximately twenty-five percent in one experiment to a high of seventy-five percent in another. Importantly, 'relativists' nonetheless drew the standard distinction between moral and conventional normative violations, viewing the former as relatively less permissible and authority contingent and more serious. Summarizing his findings, Nichols writes, "apparently people can be nonobjectivists about certain types of transgressions while still treating

such transgressions …very much in the way that objectivists treat them" (2004a, p. 25). Consistent with this view, studies (Nichols 2002) have also shown that disgust norms—at least some of which (e.g., norms regarding taboo foods) are likely viewed by subjects as culturally-relative in nature (see discussion of disgust norms in Section III)—are treated as nonconventional along the standard dimensions. Thus, the common interpretation of the moral/conventional distinction may not hold: from the mere fact that normal subjects view moral norms in nonconventional terms we cannot conclude that they are objectivists.

While the implications of Nichols' (2002b) data set should not be overstated, it appears, based on this research, that many young adults are ethical relativists, according to our definition (see Footnote 2 above). In each of the five studies, a substantial proportion of subjects denied that there is a 'fact of the matter' to settle moral disputes. This finding indicates that these individuals are not moral objectivists, since endorsing moral objectivism would seem to require a belief in ethical 'facts of the matter.' There is, of course, a possibility that additional measures could provide conflicting data. Perhaps these apparent relativists are merely confused or have objectivist moral commitments that are not tracked by the measure utilized in Nichols' (2002b) studies. At this early point in the investigation, however, the finding that so many young adults offer a relativist-style response to this apparent measure of objectivist belief poses a significant challenge to the psychological thesis at the foundation of moral error theory—i.e., that we are inherently ethical objectivists.

Perhaps the strongest empirical support for this psychological claim stems from two small studies (n=32, for both studies combined) conducted by Nichols and Folds-

Bennett (2003) involving children aged four to six. This research uncovered preliminary evidence that members of this age group may view moral properties in realist terms. In these two studies, children were asked if actions or items have a specified property (e.g., "are grapes yummy?"). If they assented, they were asked whether or not the given property relation is generalizable and preference-dependent. Generalizability questions focused on whether the property relation existed prior to the existence of human beings (e.g. "were roses beautiful way back then?"). Preference-dependence was measured by first telling the subjects that some people do not believe that the given property relation holds (e.g., "some people do not believe that pulling another child's hair is bad"), and then asking if the property relation applies "[only] for some people, or for real." Responses to the three general categories of property relations—'taste' relations ("yummy" or "icky", "fun" or boring"), beauty relations ("beautiful" or "ugly"), moral relations ("good" or "bad")--were sampled in these two studies, with the first experiment focusing on positive properties and the second on negative properties. In both studies, children tended to view 'taste' properties as preference-dependent, and were significantly less likely to judge moral properties and beautiful properties in this way (i.e., viewing them as good or beautiful "for real"). With regard to the generalizability measure, in the first study, there was no significant difference along the three property categories, with subjects tending to judge each property type as generalizable. In the second study, however, children were more likely to view moral properties as generalizable.

As Nichols' interprets them, these findings indicate that young children are moral realists, since these children tend to view ethical properties, as opposed to 'taste, properties, as preference-independent. "These results indicate that children are indeed

moral objectivists. They seem to regard moral properties as real and independent of both conventions and preferences" (2002a, p. 176). While this is a plausible interpretation of the data, more studies with larger sample sizes need to be conducted before this conclusion should be accepted. Indeed, Nichols, himself, notes, "[these findings] do not exclude the possibility that children regard moral properties as response-dependent in some other sense. The experiments probe only a simple kind of response-dependence" (2004a, p. 175). Indeed, the fact that a substantial proportion (roughly 34%) of the children sampled did not clearly identify moral properties as preference-independent—and that preference independence was not reliably correlated with generalizability—raises further questions about the implications of these findings.

Nichols concludes, however, based on this evidence, that moral objectivism is our default setting, which can occasionally be overridden. He writes, "at this early stage in the empirical exploration of intuitions about moral objectivity, the view of moral objectivity as a defeasible setting on commonsense is sufficiently promising to merit provisional adoption" (Nichols, 2004a, p. 177). In my view, there is insufficient evidence to warrant even this mitigated inference. The strongest experimental evidence in support of this claim comes from Nichols and Folds-Bennet's findings (2005) indicating that young children may have realist leanings. As noted above, however, there are questions about the reliability of the measures utilized in this research and issues involving sample size. Moreover, there is no direct experimental evidence that I am aware of indicating that a tendency to view moral norms as preference-independent typically carries over into young adulthood and beyond. In fact, there is experimental evidence to the contrary, i.e., the findings (Nichols, 2002b) outlined above that substantial proportions of college

students deny that there are moral 'facts of the matter.' Regardless, even if we were to accept Nichols' claim that we possess a natural, yet defeasible belief in moral objectivism, this would still be a problem for moral error theory. If this commitment to moral realism is easily defeasible, as the evidence suggests, then for many individuals moral error theory would not apply, since their judgments would no longer include an erroneous objectivist commitment.

In summary, moral error theorists' central claim that we are naturally committed to ethical objectivism currently lacks empirical support. As outlined above, there is a very limited range of experimental studies directly addressing this question, and the results are far from conclusive. In general, cross-cultural research, utilizing more precise measures of objectivist versus relativist belief patterns, would seem to be required in order to establish the strong claim made by moral error theorists. The moral/conventional studies that have typically been cited in support of this psychological thesis—i.e., that we are inherently ethical objectivists—falls short. Indeed, as I will argue in the next section, the finding that normal subjects tend to view moral norms as less permissible and authority-contingent and more serious than conventional norms only establishes a weaker thesis that we naturally invest moral norms with practical clout—a tendency which appears to be affectively-based. Hence, given the current range of evidence, moral error theory should be rejected on the basis of its dubious psychological foundations.

III. Practical Clout

It was argued above that moral error theory rests on an unjustified psychological assumption. The claim that we are inherently moral realists lacks empirical support. But

why have moral error theorists assumed otherwise? I believe their mistake stems from a misinterpretation of an otherwise astute psychological observation: we naturally invest morality with 'practical clout.' That is, we appear to impute a distinctive practical authority to moral norms, such that they are generally viewed as more authoritative than nonmoral norms and afforded greater weight in practical deliberation, making it more likely that moral considerations will trump nonmoral considerations in everyday decision-making. When we are considering conflicting courses of actions, 'morally required' options tend to feel more obligating-- weighty and inescapable-- than (most) nonmoral options[8], inclining us towards the former. Although failing to establish that we are inherently moral realists, the moral/conventional studies—showing that normal subjects tend to view moral normative violations as less permissible and authority-contingent and more serious than conventional ones—provide strong empirical support for the claim we naturally invest morality with practical clout. It seems that moral error theorists, in turn, have mistaken this for a belief in moral objectivism.

Indeed, the available evidence suggests that our natural tendency to invest moral norms with practical clout is rooted in emotion, not a belief in moral realism. As outlined in Chapter 1, experimental evidence indicates that affective input directly influences moral judgment in a variety of ways. Recall, for instance, that Damasio et al. (2007) found that subjects with damage to their prefrontal cortex, an area of the brain linked to emotional processing, offer aberrant evaluations of the Footbridge trolley problem scenario (i.e., is it morally permissible for someone to *push* an innocent bystander in front of a train to save five lives?), tending towards consequentialist judgments. In an

[8] As described below, disgust norms may be one type of nonmoral norm that carries a comparable degree of practical clout.

especially relevant study for our purposes, Haidt and Wheatley (2005) discovered that eliciting disgust-reaction in subjects causes them to judge moral violations as more serious, and can even lead subjects to judge that a morally-neutral scenario[9] involves an ethical violation. Hence, there is experimental evidence showing that emotion can shape our ethical assessments in various ways, including our perception of the seriousness of a violation.

Nichols' (2002) study of disgust norms, however, provides the strongest support for the thesis that emotion causes us to view ethical norms as more practically binding than conventional ones. As touched upon in the previous section, Nichols found that, just as in the case of moral norms, subjects judged violations of disgust norms—another class of norms linked to intense emotion—to be less permissible and authority contingent and more serious than conventional norms. Hence, it appears that we also naturally attribute practical clout to disgust norms. We afford disgust norms greater weight than conventional norms in practical deliberation, increasing the likelihood that the former will trump the latter. Based on his moral/conventional research, Nichols (2004a) groups moral norms and disgust norms together under the heading of "sentimental rules," distinguishing them from conventional rules, which generally lack a connection to intense affect. He concludes, "affective response infuses [moral] norms with a special nonconventional status, and this status seems to be shared by other Sentimental Rules, like norms prohibiting disgusting behavior" (2004a, p. 29). While I think that both moral norms and disgust norms naturally possess practical clout—due to similar underlying

[9] The 'morally-neutral' scenario utilized in this study was presented to subjects as follows: "Dan is a student council representative at his school. This semester he is in charge of scheduling discussions about academic issues. He often picks topics that appeal to both professors and students in order to stimulate discussion." Haidt and Bjorklund (2008, pg. 199) report that one-third of the subjects in the disgust-induction condition rated Dan's actions as "somewhat morally wrong."

affective mechanisms—there may be some difference in degree. For example, it seems plausible that moral norms may typically have *more* practical clout than disgust norms. This would be a good question for further research. For the time being, it is safe to conclude that the distinction between moral norms and disgust norms is more blurred than is commonly supposed--a thesis consistent with Haidt et al.'s research (2006) showing that in many cultures disgust norms are treated as moral in character. More importantly, for our purposes, Nichols' disgust norm research provides strong analogical evidence that our psychological tendency to impute practical clout to moral norms is affectively-based.

Of course, the finding that this natural proclivity is linked to emotion does not rule out the possibility that it also depends on an objectivist belief pattern. Perhaps our strong emotional responses to moral normative violations stem from a natural belief in moral realism, and without this belief we would no longer attribute practical clout to ethical norms? While Nichols (2004a, pg. 194) makes the plausible claim that "adults do not regard disgusting violations as objectively wrong"—the implication being that our tendency to invest disgust (and, by extension, moral) norms with practical clout is not based on an objectivist belief pattern—it is not clear that this assumption is correct. It is possible that we naturally believe that some things (e.g., spitting in one's drink) are objectively disgusting, irrespective of our subjective preferences.[10]

Nichols' moral relativist studies (2004b) provide the most compelling evidence that our tendency to impute practical clout to moral norms is not based on a belief in moral realism. As noted above, despite their ostensible rejection of moral objectivism,

[10] It is also plausible that we naturally view some things as objectively disgusting and other things (e.g., taboo foods) as culturally-relative.

ethical relativists still viewed moral normative violations as less permissible and authority contingent and more serious than conventional ones. In light of this finding, Nichols concludes, "spurning objectivity by no means eradicates the power and authority of [moral] norms…The emotions that make moral judgment distinctive continue to burn... (2004a, p. 198)." In evaluating the practical implications of evolutionary moral error theory, Ruse and Joyce draw a similar conclusion to Nichols'. Ruse underscores, "there is no question of simply breaking from morality if we so wish. Even though we have insight into our biological nature, it is still *our* biological nature" (1984, p. 104); while Joyce writes, "it is not clear what impact an epistemic ban on moral belief would have on the status of moral *emotion*...Human emotion is a much more peculiar affair than we usually think, and there are many means of influencing practical choice" (2006, p. 225-7). Indeed, Ruse and Joyce both appear to recognize that we have a natural tendency to attribute practical clout to moral norms, a proclivity which seems to be affectively-based. Their chief error is presuming that this tendency is psychologically linked to a belief in moral realism, a thesis which lacks empirical support.

The evidence that our tendency to impute practical clout to ethical norms is not dependent on a belief in moral realism should mollify concerns about the practical consequences of ethical relativism. Beginning with Plato, moral realist philosophers have long feared that the adoption of an ethical relativist belief system could lead to an upsurge of immoral behavior, since relativists will (supposedly) be less likely to attribute practical clout to moral norms. In modern times, this worry has often been linked to a concern about the loss of religious belief, which has traditionally bolstered moral realism. Dostoevsky indelibly portrayed this view through his Ivan Karamazov character, who

infamously concludes, “since god is dead, everything is permitted.” Addressing this general worry, Nichols writes, “would a rejection of moral realism engender rampant nihilism?...People worry that the abandonment of moral objectivity threatens to unravel the moral tissue of society” (2004a, p. 190). Our findings should allay some of these fears. Again, it appears that relativists view moral norms in much the same way as realists, attributing a binding force to this class of norms that affords them an extra degree clout in practical decision-making: the reason being that this natural tendency is affectively-based and apparently not dependent on a belief in moral realism. Hence, there is little reason to assume that ethical relativists will be more likely than their objectivist counterparts to flout moral norms.

Indeed, it seems that Hume was right. Philosophical reflection undercuts moral realism, but the practical implications of this understanding are limited. Although an evolutionary genealogy of our moral sensibility argues against ethical objectivism, our evolved proclivities remain largely intact. While there is insufficient evidence to conclude that we are inherently ethical realists—which undermines moral error theory, challenging its central psychological claim—the moral/conventional studies make one thing clear: we naturally invest morality with practical clout on the basis of how we feel, and this moral feeling seems to be immune to philosophical doubt. As aptly stated by Hume in the opening quote of this chapter, it appears that philosophical arguments against moral realism have “little or no influence on practice… [since] nothing can be more real, or concern us more than our sentiments..[and] no more can be requisite to the regulation of our conduct and behavior” (1964, p. 469).

Chapter 4 References

Blair, R. (1995). A cognitive-developmental approach tot morality: Investigating the psychopath. *Cognition* 57: 1-29.

Boyd, R. (1997). How to be a Moral Realist. In *Moral Discourse and Practice: Some Philosophical Approaches,* eds. Darwall, S., Gibbard, A., & Railton, P. New York: Oxford UP.

Damasio, A., Adolphs, R., Tranel, D., & Koenigs, M. (2007). *Get full reference.

Darwall, S. (1998). *Philosophical Ethics.* Boulder: Westview Press.

Haidt, J., Kollers, S., & Dias, M. (1993). Affect, culture or morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology* 65: 613-28.

Hume, D. [1793] (1964). *A Treatise of Human Nature.* Oxford: Clarendon Press.

Greene, J. (2008). The secret joke of Kant's soul. In *Moral Psychology, Volume 3* ed. W. Sinnott-Armstrong. Cambridge, Mass.: MIT Press.

Joyce, R. (2006). *The Evolution of Morality.* Cambridge, Massachusetts: MIT Press.

Kant, I. [1785] (1964). *Groundwork of the Metaphysics of Morals,* Trans. H.J. Patton. New York: Harper Torchbooks.

Mackie, J. (1977). *Ethics: Inventing Right or Wrong.* London: Penguin.

Nagel, T. (19846). *The View from Nowhere.* Oxford: Oxford UP.

Nichols, S. (2002). Norms with feeling towards a psychological account of moral judgment. *Cognition* 84: 221-236.

Nichols, S., & T. Folds Bennett. (2003). Are children moral objectivists? Children's judgments about moral and response-dependent properties. *Cognition* 90: B23-32.

Nichols, S. (2004a). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford UP.

Nichols, S. (2004b). After objectivity: an empirical study of moral judgment. *Philosophical Psychology* 17.

Railton, P. (1997). Moral Realism. In *Moral Discourse and Practice: Some Philosophical Approaches,* eds. Darwall, S., Gibbard, A., & Railton, P. New York: Oxford UP.

Ruse, M. (1986). Evolutionary ethics: a phoenix arisen. *Zygon* 21: 95-112.

Shafer-Landau, R. (2003). *Moral Realism: A Defence*. Oxford: Oxford UP.

Singer, P. (2006). Morality, reason, and the rights of animals. In *Primates and Philosophers,* ed. F. de Waal. Princeton: Princeton UP, 98-119.

Smetana, J. (1993). Understanding of social rules. In *The Development of Social Cognition: The Child as Pscyhologist,* ed. M. Bennett, 111-141. New York: Guilford Press.

Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies* 127: 109-166.

Sturgeon, N. (1985). Moral Explanations. In *Morality, Reason and Truth,* eds. D. Copp & D. Zimmerman. Totowa: Rowman and Allanheld.

Stich, S., & J. Weinberg. (2001). Jackson's empirical assumptions. *Philosophy and Phenomenological Research* 62: 637-643.

Turiel, E. (1983). *The development of social knowledge: morality and convention*. Cambridge: Cambridge UP.

Wheatley, T., & J. Haidt. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science* 16: 780-4.

5. Moral Justification Naturalized: The Method of Wide Reflective Equilibrium

"Since the history of moral philosophy shows that the notion of moral truth is problematical, we can suspend consideration of it until we have a deeper understanding of moral conceptions…In order to do this, one tries to find a scheme of principles that match people's considered judgments and general convictions in reflective equilibrium…In the [role of moral theorist], we are investigating an aspect of human psychology, the structure of our moral sensibility."

--John Rawls (1975)

In the quote above, John Rawls outlines a naturalistic approach to moral justification that accords nicely with the general orientation of my project, which has focused on empirical research regarding the intuitive and affective foundations of ethical cognition and some of the metaethical implications. Here we have a justificatory method that affords moral intuitions and feelings, as well as scientific considerations, *normative* weight. According to the standard conception in the tradition, one of the primary goals of normative ethical theorizing is to identify a set of *justified* moral judgments and principles. People espouse a wide variety of ethical views, and a central question for normative ethics is which of these claims are worthy of endorsement. Over the years, ethicists have espoused divergent approaches to moral justification. While moral philosophers typically agree that theory justification requires a special type of reasoning procedure, several different models have been proposed. Most of these approaches, however, can be classified as versions of Foundationalism. As described in Section I, foundationalist approaches seek to justify ethical judgments and principles based on supposedly self-evident moral truths. In contrast, the method of wide reflective equilibrium, which is based on an alternative, coherentist vision of moral justification, avoids any appeal to objective ethical truths. According to this approach, judgments and principles are endorsed based on their overall coherence within a larger network of

considerations. We begin by incorporating those intuitive ethical judgments and principles of which we are most confident, as well as relevant metaethical and scientific theories, and then work back and forth between these considerations, revising or rejecting old views and introducing new ones, until an 'equilibrium point' is reached in which our beliefs are maximally coherent. The set of judgments and principles endorsed in wide reflective equilibrium are deemed justified. Section I delineates the general structure of this normative reasoning method, while further distinguishing it from foundationalist approaches.

In light of this broad outline of the method of wide reflective equilibrium, sections II and III focus on two distinguishing, naturalistic features of this coherentist approach to moral justification. Section II clarifies how this method affords our gut feelings normative weight, while the last section underscores the important justificatory force of relevant scientific theories. To my awareness, in the long history of moral philosophy, there has not been another approach to ethical justification incorporating both of these naturalistic features. Indeed, each, taken alone, is quite uncommon. Other than the Classical Intuitionists, such as H.A. Prichard (1912) and W.D. Ross (1939), very few philosophers have argued that our moral intuitions have justificatory force; and none that I am aware of have developed a normative reasoning procedure that attributes justificatory force to the moral feeling linked to our intuitions. Section II fleshes out how the method of wide reflective incorporates emotional coherence considerations in the overall calculation that determines belief acceptability, based on Paul Thagard's "HOTCO" connectionist models of coherence. This naturalistic feature of the method is a response to the central role of emotion in belief acceptance and motivation. Empirical

evidence indicates that the reason we tend to be so committed to our intuitive beliefs is the strong emotion to which they are connected. We have a strong psychological proclivity to remain committed to our gut feelings, and it is unrealistic to assume that we could accept and act upon a set of normative beliefs that do not largely cohere with them. By affording our moral feelings justificatory force, the method of wide reflective equilibrium ensures a psychologically realistic result.

This method is also naturalistic in the sense that it affords empirical findings and theories of the sort incorporated in this project (e.g., research concerning our moral psychology, cross-cultural ethical practices, the evolution of morality, etc.) justificatory force. This is yet another distinctive feature, as the vast majority of approaches to moral justification in the tradition fail to attribute normative weight to scientific considerations. Section III outlines how empirical theories may carry substantial justificatory force in the method of wide reflective equilibrium, in response to Paul Thagard's (2010) misguided[11] criticism that this coherentist approach lacks empirical grounding. Finally, it is argued that the Neo-Aristotelian justificatory method endorsed by Thagard, which is another approach to moral justification affording normative weight to scientific findings, suffers from the same characteristic weaknesses as other foundationalist methods: a failure to satisfactorily establish its normative foundations and related problems of under-determination. Accordingly, I contend that the method of wide reflective is a superior naturalistic approach. Finally, in the last section of this chapter, I address the future of normative ethics in light of the modern 'science of morality.'

[11] Misguided and ironic, given Thagard's earlier work, 2002, expounding this coherentist method. See Section II below.

I. The Method of Wide Reflective Equilibrium

While much has been written about Rawls' political philosophy and theory as famously articulated in *A Theory of Justice* (1999), very few moral theorists have specifically addressed his justificatory method, and most of these commentaries have been critical. Commentators, such as R.M Hare (1989), Richard Brandt (1998) and Paul Thagard (2010), argue that Rawls' coherentist approach to moral justification is little more than a sophisticated version of Classical Intuitionism, which is a foundationalist method. Accordingly, one of the aims in this chapter is to properly distinguish the method of wide reflective equilibrium from Classical Intuitionism, in order to show that this common criticism misses the mark. This section sets the stage by outlining the general structure of this coherentist justificatory procedure, as compared to foundationalist methods, broadly construed. My characterizations of wide reflective equilibrium in this chapter is based, in large part, on the description provided by Norman Daniels (1979, 1980, 1996), who has offered the most systematic and forceful defense of this approach in the literature. Nonetheless, there are features of this method, such as the two naturalistic ones mentioned above, which are not properly emphasized by Daniels. The general goal here is to illuminate these distinctive characteristics in order to clear up confusion and further bolster this method.

As emphasized above, the method of wide reflective equilibrium offers an alternative to foundationalist approaches to ethical justification, which have traditionally been the most popular, despite their characteristic flaw. Broadly speaking, an approach to moral justification involves two main, closely interrelated facets: (1) a conception of what counts as justification, and (2) a corresponding procedure for determining which

ethical judgments and principles are worthy of endorsement. One general type of approach is Foundationalism. Foundationalist views assume that there are normative considerations, typically a set of moral judgments or principles, which are self-evidently true and justified. These methods characteristically rely on a moral realist conception of truth, which was defined in Chapter Four as a general position stipulating that "there are 'independent' ethical truths or facts that apply irrespective of our subjective preferences." According to the standard foundationalist conception, ethical judgments and principles are justified when they are objectively true, and we can determine which beliefs are objectively true based on foundational beliefs that are self-evident. Hence, the general method recommended by foundationalist thinkers is to first identify these epistemological Archimedean points and then evaluate other judgments and principles on this basis. For instance, Kant's (1964) normative theory starts with a fundamental principle—the Categorical Imperative ("I am never to act otherwise than so *that I could also will that my maxim become a universal law"*). The Categorical Imperative is then used as a standard for evaluating judgments and "maxims" concerning our moral obligations. Only maxims that pass this test are deemed justified. Kant's is a 'top-down' foundationalist approach, utilizing a broad principle to justify discrete moral rules. Other foundationalist approaches, such as the one characteristic of Classical Intuitionism (more on this in the next section), work from the 'bottom-up'; starting with an epistemologically-privileged set of judgments or intuitions, from which more general principles are formulated through induction.

Regardless of whether they are 'top-down' or 'bottom-up,' foundationalist approaches typically suffer from the same weakness: a failure to satisfactorily establish

the legitimacy of their normative foundations. More often than not, foundationalist thinkers merely assert or assume the self-evidence of their favored foundational principles, rather than providing a suitable justification. This is hardly satisfactory as virtually every foundational judgment or principle proposed by one philosopher has been challenged by others. Moreover, these foundationalist approaches are typically underdetermining. It seems highly dubious that a small set of normative Archimedean points could serve to justify a full range of moral beliefs given the great diversity of moral problems and questions we face. As highlighted below in Section III, merely identifying foundational principles or judgments is not enough. We also require some means of ordering and ranking (i.e., by ethical importance) these principles when they conflict. Moreover, it can be difficult to apply these broad principles to concrete cases. For example, we may know that it is wrong to be greedy, but be unsure as to whether or not Bill Gates is guilty in this regard. Answering these difficult questions requires a more nuanced justificatory reasoning procedure than the foundationalist picture suggests.

The main weakness of foundationalist approaches, however, is their characteristic appeal to moral realism. Again, according to this conventional view of ethical justification, moral beliefs are justified when they are objectively true, and the veracity of some beliefs (i.e., those which are to serve as justificatory Archimedean points) is supposedly indubitable. As argued in Chapter Four, however, a Darwinian genealogy of morality poses a significant epistemological challenge to moral realism. There are strong evolutionary grounds for avoiding the postulation of objective moral truths—and so the conventional foundationalist conception of moral justification is not viable. Just to be clear, the primary problem for Foundationalism is not the assumption that some basic

moral principles (e.g., 'the gratuitous infliction of pain is wrong") are likely justified. As will be addressed below, the method of wide reflective equilibrium incorporates a small number of core beliefs as provisional justificatory fixed points, which are highly unlikely to be deemed unjustified. The issue is *how* these core beliefs (and other moral beliefs) are ultimately endorsed. In contrast to foundationalist approaches, the method of wide reflective equilibrium avoids any appeal to supposedly self-evident truths. Daniels underscores, "wide reflective equilibrium embodies coherence constraints on theory acceptance or justification, not on truth" (1979, pg. 277). In the passage cited at the opening of this chapter, Rawls emphasizes a similar point, noting that this justificatory procedure "does not presuppose the existence of objective moral truths" (pg. 8). According to this approach, all moral beliefs, including provisional fixed points, must be justified in the same way: by demonstrating how they cohere with the set ultimately endorsed in wide reflective equilibrium. Hence, this coherentist approach to moral justification overcomes the major limitation of traditional, foundationalist alternatives by avoiding any appeal to ethical realism, which is an empirically suspect view.

Daniels (1979) summarizes the method of wide reflective equilibrium in the following way,

> The method of [wide] reflective equilibrium is an attempt to produce *coherence* in an ordered triple of sets of beliefs held by a particular person, namely, (a) a set of considered moral judgments, (b) a set of moral principles, and (c) a set of relevant background theories. We begin by collecting the person's initial moral judgments and filter them to include only those of which he is relatively confident…We then propose alternative sets of moral principles that have varying degrees of 'fit' with the moral judgments…[Then], we advance philosophical arguments intended to bring out the relative strengths and weaknesses of the alternate sets of principles (or competing moral conceptions). These arguments can be construed as inferences from some set of relevant background theories (I use the term loosely)…We can imagine the agent working back and forth, making adjustments to his considered judgments, his moral principles, and his background theories. In this way he arrives at an equilibrium point that consists of the ordered triple (a), (b), (c) (pg. 257).

As noted above, an approach to moral justification involves two facets, a conception of what counts as justification and a corresponding procedure for determining which moral judgments and principles are worthy of endorsement. The passage above focuses on the methodological side of wide reflective equilibrium, which, in turn, rest on a particular conception of ethical justification. According to this approach, moral judgments and principles are justified based on their overall coherence within a larger network of beliefs, including candidate (a) judgments, (b) principles, and (c) background theories. Within this larger set, some beliefs will support each other, while others will conflict. Normative justification is determined by a reasoning procedure that works to sort through this initial

set of considerations (a, b, c) in order to produce a set of beliefs that are maximally coherent. The ultimate goal is to arrive at wide reflective equilibrium, a point at which our (a) ethical judgments, (b) principles and (c) background theories are in harmony with each other. Based on this conception, the moral judgments and principles endorsed in reflective equilibrium are tentatively justified.

This coherentist approach incorporates three types of considerations or beliefs, a set of (a) moral judgments, (b) ethical principles and (c) "background theories." According to the standard conception in the moral tradition, ethical judgments address *particular* cases or moral situations (e.g., 'Jane was wrong to steal a book from Howard'), whereas ethical principles concern more *general* obligations or rules for action (e.g., "stealing is wrong"). Moral principles, in turn, can vary in complexity, from the simple example above to more nuanced formulations (e.g., 'stealing is wrong, except under the following conditions…'). Some normative theories incorporate a hierarchy of principles, in which some are derived from others. For example, Kant (1964) deduces general maxims (e.g., 'one must not tell a lie') from a more fundamental principle, the Categorical Imperative. With regard to level (c) "background theories," Daniels indicates that there are two main types, metaethical and scientific. As touched upon at the beginning of Chapter Three, metaethics concerns a variety of issues, including the nature of ethical truth or justification, judgment and learning, the function or purpose of morality, as well as the meaning of ethical terms, like 'goodness,' 'right' and 'duty.' Although Daniels (1980) does not explicitly identify them in this way, most of the examples he provides of (c) background theories are clearly metaethical in nature. He cites examples, such as a theory of the person, a theory of procedural justice and a theory

of the role of morality in society, drawn from Rawls' theory of justice. For instance, as part of his conception of the person, Rawls postulated that we are rationally, self-interested individuals. Rawls' theory of the role of morality in society focuses on social justice. According to this conception, the primary function of morality, at least from a societal perspective, is furnishing principles of justice to govern broader institutional structures and practices (e.g., the redistribution of wealth). These theories qualify as metaethical conceptions, based on the broad definition above.

The other main type of level (c) consideration identified by Daniels—i.e., scientific findings and research—is closely tied to metaethics. As demonstrated throughout this project, empirical findings can have important implications for traditional metaethical issues, e.g., debates concerning the nature of moral truth and judgment. Given this connection, it is not surprising that Daniels incorporates both metaethical theories and scientific research at level (c) of this coherentist approach to justification. He writes,

> [The] marshalling of the broadest evidence and critical scrutiny is the attraction of *wide* as opposed to *narrow* reflective equilibrium. We not only must work back and forth between principles and judgments about particular cases, the process that characterizes narrow equilibrium, but we must bring to bear all theoretical considerations that have relevance to the acceptability of the principles as well as the particular judgments. These theoretical considerations may be empirical or they may be moral. One task of ethical theory, then, is to show how work in the social sciences, for example, has a bearing on moral considerations (1996, pg. 6).

In this passage, Daniels draws an important distinction between "narrow" and "wide" versions of reflective equilibrium, a contrast which will receive greater attention in Section III below. Daniels repeatedly emphasizes the inadequacy of narrow reflective equilibrium, which only involves two types of considerations, (a) judgments and (b) principles. He underscores that this version of reflective equilibrium, as opposed to the wide variety he endorses, lacks the resources to substantially challenge our (a) and (b) intuitions—a characteristic weakness of Classical Intuitionism as well. He contends that the method of wide reflective equilibrium overcomes this shortcoming by incorporating level (c) considerations, of which there are two main types, "moral" and "empirical"—i.e., metaethical and scientific. Again, with regard to the latter, in the long history of moral philosophy there have been very few approaches to ethical justification that attribute normative weight to scientific considerations. Accordingly, Daniels emphasizes that this coherentist approach "expand(s) the kinds of considerations that count as evidence for or against our moral views at all levels of generality" (1996, pg. 6). In section III, this naturalistic feature of wide reflective equilibrium will be addressed in greater detail, with specific examples of how scientific considerations can lead to the revision and endorsement of beliefs at all three levels (a, b, and c). It will be shown that although scientific theories are not, themselves, justified through the method of wide reflective equilibrium, these theories play an important justificatory role in this coherentist method.

Returning back to the basic structure of the method of wide reflective equilibrium, Daniels also recommends a specific sequence for incorporating the three types of considerations (a, b, and c) into this justificatory procedure, a sequence which I do not

endorse. He suggests that we begin by admitting a set of (a) moral judgments into normative consideration. Then, we are supposed to formulate (b) moral principles in an effort to systematize these (a) judgments. Finally, we can introduce relevant (c) background theories to help decide between competing (b) principles. Once all these considerations are in play, we then work towards an equilibrium point by adjusting our beliefs at all three levels. As noted above, Rawls indicates that one of the primary goals of the method of wide reflective equilibrium is to "investigate" our moral psychology by incorporating our pre-reflective ethical intuitions and convictions into the justificatory reasoning process; so that we can evaluate which beliefs are worthy of endorsement. Accordingly, Daniels proposes what I am calling the "confidence criterion," a standard for determining which level (a) judgments will be initially admitted into normative consideration. He writes, "we begin by collecting the person's initial moral judgments and filter them to include only those of which he is relatively confident… (1979, pg. 258)." As emphasized in the next section, the main implication of the confidence criterion for the method of wide reflective equilibrium is that many of the beliefs initially introduced as normative considerations will be 'hot' moral intuitions; since these are the beliefs of which we are typically most confident. In this way, the confidence criterion accords with Rawls' goal of normatively evaluating our moral sensibility. If we want to test our moral sensibility, then we need a mechanism for incorporating our intuitions and gut feelings into this coherentist justificatory procedure. The confidence criterion serves this function.

It seems, however, that the confidence criterion should be applied at all three levels of this coherentist approach, and not just at level (a), as suggested by Daniels. As

demonstrated in Chapter 1, in addition to those pertaining to particular moral cases and scenarios, we also have strong moral intuitions relating to general moral principles or obligations (e.g., 'it is wrong to lie'). Furthermore, the studies of experimental philosophers, such as Shaun Nichols (2002) and Eddy Nahmias (2007), indicate that we have metaethical intuitions, e.g., with regard to the meaning of ethical terms, the sources of moral responsibility, etc. Given the goal to investigate our ethical sensibility, it seems that these (b and c) intuitions should be initially incorporated in the method of wide reflective equilibrium as well. The sequence Daniels outlines--i.e., start with (a) judgments (filtered based on the confidence criterion), then formulate (b) principles on this basis, and, finally, consider (c) background theories--would only incorporate moral intuitions at level (a). Hence, I think the initial sequence he recommends ought to be abandoned, and we should instead begin this justificatory procedure by incorporating intuitions at all three levels (a, b, c), based on the confidence criterion. Indeed, this revised starting procedure would also be more in the general spirit of this coherentist approach to justification, which is not supposed to epistemologically privilege any particular type (a, b, or c) of belief. According to this method, once beliefs, at any level, are admitted into the justificatory process, their normative weight is determined solely by coherence considerations. In principle, all three *types* of moral beliefs (a, b, c) are initially afforded the same epistemological status, i.e., *as normative considerations, pending justification*.[12] Hence, it is strange that Daniels suggests an initial sequence that

[12] I am not including scientific considerations under the category of 'normative considerations, pending justification', since as noted above they are not subject to justification via this normative reasoning procedure. On this account, however, the other type of level 'c' consideration, metaethical beliefs, falls under this category. In this chapter, 'moral beliefs' or 'normative considerations' refers only to level 'a' and 'b' beliefs as well as level 'c' metaethical considerations. While the scientific considerations incorporated by this method are normatively *relevant*, they are not normative *considerations*, per se.

would, in effect, epistemologically privilege level (a) beliefs, such that they would determine the (b) principles and (c) background theories that are initially incorporated by this method. Again, this would be an undesirable result given that we want this coherentist approach to evaluate our moral sensibility, which is comprised of intuitions at all three levels.

Let us now revisit the question of how the method of wide reflective equilibrium differs from foundationalist approaches to moral justification, broadly conceived. This coherentist approach incorporates three types (a, b, c) of beliefs. We begin by collecting those moral beliefs at all three levels of which we are most confident. In principle, every normative consideration in this initial set is subject to revision, although some of these beliefs will be relatively more central and thus less likely to be abandoned (more on this below). Contrast this to foundationalist approaches, in which some normative considerations are assumed to be justified *right from the start*. These fixed, "Archimedean" points—the justification of which is supposedly beyond doubt--serve as anchors for the endorsement of other moral beliefs. As emphasized earlier, the main problem for foundationalist approaches is their reliance on supposedly 'self-evident' moral truths. The method of wide reflective equilibrium avoids this difficulty. Daniels emphasizes, "no considered moral [beliefs] at any level are taken to be unrevisable, that is, *strongly foundational*; moreover, they are subject to revisionary pressures from considerations at all levels" (1980, pg. 83). According to this approach, moral justification is an ongoing process and equilibrium points are tentative, pending the incorporation of new beliefs and theories (or the reformulation of old ones).

Granted, although all normative considerations in this coherentist network are theoretically subject to revision, some are highly unlikely to be rejected. It is very hard to conceive of a wide reflective equilibrium point that would not include, for example, basic principles, such as 'the gratuitous infliction of pain is wrong' or 'helping others in need is good.'[13] These central beliefs are so deeply intertwined with the moral judgments, principles and metaethical conceptions of which we are most confident that it would almost be inconceivable for these core beliefs to not be among the set endorsed in wide reflective equilibrium. It is in this sense that Daniels refers to "provisional fixed points" in the justificatory network. He writes,

> since all considered [beliefs] are revisable, the [belief] 'it is wrong to inflict pain gratuitously on another person' is too. But we can also explain why it is so hard to imagine not accepting it…To imagine revising such a provisional fixed point we must imagine a vastly altered wide reflective equilibrium that is nevertheless much less acceptable than our own. For example, we might have to imagine persons quite unlike the persons we know (1979, pg. 267).

Even though there will be provisional fixed points in this coherentist justificatory procedure, this is based solely on coherence (the two main senses of coherence, explanatory and emotional, will be outlined below) considerations. As emphasized in the next two sections, by incorporating our moral intuitions as normative considerations, this method does not automatically justify them. While some of these beliefs will serve as provisional fixed points, others will be rejected, based on their degree of 'fit' with considerations at all three levels.

---

[13] There is, of course, substantially more room for variability concerning the justified exceptions to these general rules.

Indeed, the relative weight of normative considerations, including provisional fixed points, in the method of wide reflective equilibrium is never fixed. All of the beliefs admitted into this coherentist reasoning procedure have a degree of justificatory force. Nonetheless, the *degree* of justificatory force or normative weight carried by individual beliefs in the network varies based on their overall level of coherence. Core beliefs, which have the most support from all three levels, carry the highest degree of normative weight, such that we will be more likely to endorse those views with which they are consistent and reject those with which they are inconsistent. In comparison, beliefs with weaker support from the network carry less normative force. An especially distinctive feature of this approach--as compared to foundationalist methods, which attribute a static justificatory force to their respective foundations--is that the normative weight of any individual belief is subject to change based on changes to this dynamic network. As old beliefs are revised or rejected, and new ones are incorporated, the justificatory weights within the network will shift as well. The amount of 'shake up' in the system will, in turn, depend on the justificatory force of the newly introduced beliefs, which, again, is based solely on coherence considerations. If a new view 'fits' with a large set of existing beliefs in the network, this belief will carry relatively more normative weight than a less coherent addition. In the passage above, Daniels describes the normative reasoning procedure linked to this coherentist approach as "working back and forth, [and] making adjustments" at all three levels. What Daniels fails to make explicit is that one of the things being adjusted is the justificatory force of beliefs in the network. Indeed, this is what normative reasoning consists of according to this approach:

evaluating the degree of 'fit' among considerations in the network and revising the individual normative weights, accordingly.

This section has delineated the basic structure of the method of wide reflective equilibrium as a coherentist approach to moral justification that avoids traditional, foundationalist appeals to realist truth. According to the method endorsed here, the normative weight of moral beliefs, and whether or not they are ultimately justified, is determined solely by their level of coherence within a dynamic network. This general outline raises a central question: what exactly does 'coherence' amount to within this framework? The next section will provide a more precise account based on neural network modeling of explanatory and emotional coherence. Indeed, coherence is not based solely on 'cold' considerations (explanatory coherence). Rather, our gut feelings (emotional coherence) are also factored in as well. Indeed, this coherentist method affords our moral feelings justificatory force, which is one of its most distinguishing naturalistic features.

II. Gut Feelings and Emotional Coherence

As noted at the outset of this chapter, the method of wide reflective equilibrium is a naturalistic approach to moral justification in two main senses. This section focuses on how this method attributes justificatory force to our moral intuitions and the gut feelings that are typically linked to them. According to this approach, coherence evaluations are based, in part, on emotion. Our feelings and convictions impact the normative weight of the beliefs to which they are associated and play an important role in the coherence evaluations that ultimately determine the set of beliefs endorsed in wide reflective

equilibrium. This naturalistic feature needs to be emphasized since it has been unacknowledged in the literature, including the writings of Rawls and Daniels. Indeed, to my awareness, in the history of moral philosophy, there has not been another *developed* theory of how emotion should be afforded normative weight within a justificatory reasoning procedure. The prevailing, Rationalist approach to ethical justification, as espoused by philosophers such as Plato (1961) and Kant (1964), holds that emotion is without normative weight. Even in the Sentimentalist tradition, there has not been a clearly articulated position regarding if and how our affective biases should be afforded justificatory force. Hume (1964) famously wrote that "reason is, and ought only to be the slave of the passion" (more on Hume in Section IV below), but he never fully clarified the implications for normative theorizing. Contemporary Sentimentalists, such as Jonathon Haidt (2008) and Antonio Damasio (2000), emphasize that ethical reasoning is *influenced* by affect. This is not the same thing, however, as claiming that moral feeling has justificatory force. By contrast, as a psychologically realistic approach to moral justification, the method of wide reflective equilibrium accounts for the central role of emotion in moral judgment and motivation by affording our feelings normative weight. As compared to Classical Intuitionism, however, a method that inadvertently justifies our pre-reflective biases and feelings, the coherentist approach endorsed here attributes a substantially more modest degree of justificatory force to moral emotion.

Let us begin by focusing on the role of ethical intuition in the method of wide reflective equilibrium, which will help to clarify the way in which this approach affords moral feelings justificatory force. As touched upon in the previous section, at the initial stages of this coherentist reasoning procedure, beliefs are admitted as normative

considerations based on the confidence criterion. Since, for practical reasons, we cannot normatively evaluate all of our moral beliefs, and given the desire to "investigate" our moral sensibility, we initially admit only those beliefs of which we are most confident. Based on the empirical evidence outlined in the preceding chapters, it appears that the moral beliefs that are typically held with greatest conviction are the intuitive ones, which are characteristically 'hot.' According to the empirical account developed in this project, moral intuitions are *beliefs* or *judgments*[14] (e.g., "theft is wrong") that are issued relatively automatically, prior to conscious moral reasoning, and are typically linked to gut feelings. In Chapter 1, Jonathon Haidt's affective biases model of intuitive ethical judgment was endorsed. According to this account, very roughly stated, our intuitive judgments are *caused* by the triggering of evolved affective predispositions. Haidt proposes that we come equipped with a set of "moral modules," each of which has an affective valence and is sensitive to a particular type of ethical situation. For instance, our 'fairness' module responds to cases pertaining to norms of reciprocal exchange (e.g., failure to pay a debt). Once these modules are 'lit up', an affective response is triggered, either positive or aversive, which typically leads to a corresponding moral judgment (e.g., 'John was wrong to not pay his debt').

Regardless of whether or not Haidt's affective biases model is correct, there is overwhelming evidence that our intuitive beliefs are linked to emotional responses and the relative intensity of these affective reactions directly impacts the level of conviction with which these beliefs are held. As discussed in Chapter 1, studies of intuitive moral judgment reveal that we have a strong tendency to remain committed to our gut ethical feelings, even when we are unable to provide adequate reasons for endorsing them--a

[14] In this project, the terms "moral belief" and "moral judgment" are used interchangeably.

phenomenon that has been termed "moral dumbfounding." Accordingly, in Chapter 3, it was argued that intuitive judgments are characteristically hotter (i.e., associated with more intense affect) than assessments based on conscious moral reasoning; and, in general, the hotter the belief, the greater the level of motivational force, such that weakness of will is less likely to occur in the case of judgments that are more emotional. Finally, in Chapter 4, the practical clout of moral norms (i.e., the tendency of ethical considerations to trump nonmoral ones in practical decision-making) was attributed to the relatively stronger emotional responses to which they are commonly linked.

These empirical findings have direct implications for the method of wide reflective equilibrium, as a psychologically realistic approach to moral justification. As noted above, with regard to the confidence criterion, the implication is clear: many of the (a) and (b) beliefs, and at least some of the (c) metaethical considerations, initially admitted as *normative considerations* will be moral intuitions, since these are the beliefs of which we are typically most confident. Although Daniels is not very forthcoming regarding this feature of the method, i.e., that it includes ethical intuitions as normative considerations, he does, at times, offer an acknowledgement. For instance, he writes,

> just what role should be assigned to moral judgments or moral intuitions in the process of selecting among or justifying moral theories is a matter of ancient controversy…This old debate has taken on a modern form in the contrast between two recent proposals for solving the problem of..justification in ethics, the method of wide reflective equilibrium proposed from Rawls and the moral empiricism advocated by Brandt (1996, pg. 81).

Brandt, whose ethical view we will not delve into here, condemns the method of reflective equilibrium for its "intuitionism" (Daniels, 1996, pg. 82). Indeed, several critics of this Rawlsian approach, including R.M. Hare (1989) and Paul Thagard (2010), argue that this coherentist approach is nothing more than a sophisticated form of Classical Intuitionism. This may explain why Daniels avoids highlighting that this coherentist approach incorporates our intuitions as normative considerations.

Nonetheless, although the method of wide reflective equilibrium attributes justificatory force to moral intuitions, it does so in a way quite different from Classical Intuitionism. Classical Intuitionism is a foundational approach that epistemologically privileges ethical intuitions, treating them as justificatory Archimedean points. Proponents of this approach, such as H.A. Pritchard (1912) and W.D. Ross (1939), claim that there is a set of moral intuitions (e.g., the intuition 'that we ought to pay our debts' or 'tell the truth') that are self-evidently justified, and we can build our ethical knowledge on this basis. Providing a nice summary of this approach, Ross writes, "I suggest…that both in mathematics and in ethics we have certain crystal-clear intuitions from which we build up all that we know about the nature of numbers and the nature of duty…In the course of thinking we come to know more, but we should never come to know more if we did not *know* what we start with" (pg. 144-5). Clearly, as described in the previous section, the method of wide reflective equilibrium operates differently. The intuitions incorporated by this coherentist approach are normative considerations, pending justification. Although these intuitions are afforded justificatory force, and some will end up serving as provisional fixed points, none are automatically deemed justified prior to philosophical reflection. Daniels emphasizes, "wide reflective equilibrium does not

merely systematize some determinate set of judgments. Rather, it permits extensive revision of these moral judgments. There is no set of judgments that is held more or less fixed as there would be on a foundationalist approach (1979, pg. 266-267)." Again, according to this method, the normative weight of individual beliefs is based solely on coherence considerations, and weights can shift as revisions are made to the network. Hence, the justificatory force of any individual intuition, even those which are core, is subject to change. This is in sharp contrast to Classical Intuitionism, which attributes a static, foundational force to the moral intuitions it incorporates.

The method of wide reflective equilibrium also affords our gut feelings justificatory force in a more modest way than Classical Intuitionism. Paul Thagard's neuro-computational models of explanatory ("ECHO") and emotional coherence ("HOTCO II") provides a useful model for our purposes. Let us focus first on the ECHO model, which does not incorporate emotion in assessing coherence. This will help to clarify how affect is afforded justificatory force within the HOTCO II program, as a means of simulating the emotional coherence calculations that enter into the method of wide reflective equilibrium. In *Coherence in Thought and Action,* Thagard (2002) provides a neuro-computational model of wide reflective equilibrium, which falls short by leaving out emotional coherence. He writes,

> The term "wide reflective equilibrium" is used to describe a state in which a thinker has achieved a mutually coherent set of ethical principles, particular moral judgments, and background beliefs. But how people do and should reach [wide] reflective equilibrium has remained poorly specified. [I will] show how we can justify ethical principles and particular judgments by taking into account a wide range of coherence considerations (126).

Thagard outlines four types of 'cold' coherence calculations that enter into the method of wide reflective equilibrium: "deliberative coherence" between moral actions and goals, "deductive coherence" among moral principles and particular judgments, "explanatory coherence" between principles and judgments on the one hand, and empirical facts and hypotheses on the other, and, finally, "analogical coherence" between moral situations (e.g., abortion and euthanasia with regard to 'right to life' issues). It would be beyond our scope to focus in detail on all of these different types of coherence relationships, which have in common that coherence evaluations are based on constraint satisfaction among beliefs within a dynamic network. While it seems that all four types of coherence relationships are part of the method of wide reflective equilibrium, as Thagard suggests, we will focus only on Thagard's ECHO model of explanatory coherence, which maps on nicely to his HOTCO II program.

Roughly outlined, the ECHO model (Thagard 2008) represents beliefs as individual units, each of which has a variable activation level, ranging between -1.0 and +1.0. The activation level simulates the acceptability (or lack thereof) of a given proposition as a function of its coherence in the network. Beliefs that cohere are connected by excitatory links that spread activation symmetrically between them, and

those which are incoherent are connected by symmetric inhibitory links. Various degrees of coherence or incoherence are represented by connection weights that determine the relative strength of the excitatory or inhibitory links between elements; and the stronger the connection weight, the greater the spread of activation or inhibition. Activations are cycled throughout the network until a stable state (i.e., wide reflective equilibrium) is reached, at which time some units (i.e., those which are 'accepted') have a positive activation and others have a negative one (i.e., those which are 'rejected'). The ECHO model simulates wide reflective equilibrium as a cold reasoning process that determines the degree of fit between beliefs, irrespective of the emotion that may be linked to them and the corresponding degree of confidence with which they are held prior to philosophical reflection. According to this model, the normative weight of an individual belief corresponds to its activation level, which is determined solely by explanatory coherence with other elements in the network. At the beginning of the process, all beliefs have the same activation level (i.e., 0), and connection weights represent an estimation of the positive and negative constraints between elements. Beliefs that meet a certain activation threshold (e.g., >0) in wide reflective equilibrium are deemed justified.

Thagard's ECHO model of wide reflective equilibrium nicely captures Rawls and Daniels' cold-processing account, as well as its main weakness. Despite their commitment to providing a psychologically realistic justificatory method, both Rawls and Daniels fail to incorporate emotion into this coherentist procedure. They appear to wrongly assume that the confidence with which our intuitions are held is independent of the feelings linked to them, and so we can evaluate the coherence of these beliefs without factoring in emotional coherence. For example, in discussing the confidence criterion,

Daniels writes, "we begin by collecting the person's initial moral judgments and filter them to include only those of which he is relatively confident and which have been made under conditions conducive to avoiding errors of judgment. For example, *the person is calm* and has adequate information about the cases being judged" (258, emphasis mine). In this passage, Daniels proposes two standards in connection with the confidence criterion that appear to be at odds. He suggests that we should initially incorporate into normative consideration only those beliefs that (1) are held with conviction and (2) issued in a "calm state." From a psychological perspective, this seems like an untenable criterion, since the beliefs of which we are most confident tend to be the hottest ones (i.e., not issued in a calm state). This error provides evidence that Daniels' overlooks the psychological connection between judgments of acceptability and emotion, which may explain his endorsement of a psychologically unrealistic, cold-processing model of wide reflective equilibrium. By failing to incorporate emotional coherence, the ECHO model opens up the possibility of endorsing a set of beliefs in wide reflective equilibrium that do not sufficiently accord with our gut feelings. Given the central role of these feelings in moral judgment and motivation, this would clearly be an unacceptable result. As described above, the empirical evidence is clear: moral intuitions are characteristically linked to gut feelings, which strongly influence our judgments regarding the acceptability of these beliefs. In general, the stronger the associated emotion, the more likely we are to endorse the intuitive belief and remain committed to it. This proclivity must be accounted for by the method of wide reflective equilibrium if it is to serve as a psychologically realistic justificatory approach.

Fortunately, Thagard (2008) has provided the necessary resources. His HOTCO II model expands upon the ECHO model outlined above by incorporating emotion as part of the overall coherence calculations that determine belief acceptance. In addition to an activation level, each unit in the HOTCO II model has a valence, either positive or negative, which simulates an emotional attitude associated with a represented belief. Thagard writes, "on this theory, mental representations such as propositions and concepts have, in addition to the cognitive status of being accepted or rejected, an emotional status called a valence, which can be positive or negative depending one's emotional attitude toward the representation "(149). For example, a unit representing the belief that 'Jane stole from her neighbor' will likely be associated with a negative valence, simulating a disapproving attitude toward the act. Valences in the HOTCO II model are calculated in a similar fashion to activation levels. Valences spread throughout the network based on the same excitatory and inhibitory channels that determine activation levels, and the overall valence of a unit is a function of the valence of all the units to which it is connected, as well as their activation levels.[15] In the original HOTCO model, activation levels impact valence calculations, but not vice versa. By contrast, Thagard's HOTCO II model makes the activation level of a unit dependent in part on its valence, with positive valences enhancing activation levels and negative valences suppressing them. Thagard (2008) writes,

[15] In mathematical terms, as characterized by Thagard (2000), "the valence of a unit u^j is the sum of the results of the multiplying, for units u^i to which it is linked, the activation of u^i times the valence of u^i, times the weight of the link between u^i and u^j (pg. 174).

> In the original version of HOTCO , the valence of a unit was calculated on the basis of the activations and valences of all the units connected to it…HOTCO enabled cognitive inferences such as the ones based on explanatory coherence to influence emotional judgments, but did not allow emotional judgments to bias cognitive inferences…Accordingly, I have altered HOTCO to allow a kind of biasing of activations by valences…The activation of [units in HOTCO II] is a function not only of the activation input to them but also of the valence input to them that they receive from the valence unit (148).

Although the manner in which it incorporates emotion needs to be slightly adjusted, HOTCO II provides a far more psychologically realistic model of the method of wide reflective equilibrium than ECHO. By contrast to the cold-processing ECHO model, HOTCO II incorporates emotional coherence considerations as well. In HOTCO II, the acceptability of a belief is partially determined by the feeling to which it is connected.

There is one facet of HOTCO II, however, that needs to be amended for our purposes. On the model sketched by Thagard, beliefs linked to positive emotions are more likely to be accepted than those associated with negative emotions. This does not accurately capture the relationship between moral feeling and conviction revealed by the empirical evidence cited above. This research indicates that it is the relative intensity of the emotion linked to moral intuitions that directly impacts our judgments regarding the acceptability of these beliefs, irrespective of whether this emotion is positive or negative. Indeed, it seems that many of our core convictions are linked to negative feelings. For example, the judgment that 'I ought not steal from my neighbor' may be strongly supported by an aversive feeling associated with the mental representation of this action

(e.g., I imagine how my neighbor would feel if I stole from her, and the aversive feeling elicited by this imaginative representation bolsters my conviction that I ought to refrain from this harmful action). It appears that many of our gut moral feelings operate in a similar fashion, i.e., as aversive 'alarm bells' signaling that one should refrain from the represented action. This is not to claim, however, that positive emotions do not also play an important role in intuitive ethical judgment. In some cases, it may be a pleasant gut feeling that leads to moral conviction. For example, I might judge that I ought to give to a charitable cause because the representation of this action generates feelings of warmth. It also seems plausible that, at least on some occasions, our core ethical intuitions are bolstered by both positive and negative feelings. For instance, my decision to give money to a charitable cause may be reinforced by both a positive feeling and a negative one (e.g., the pity I feel for those in need). In general, it seems that Thagard's HOTCO II model, which only allows for positive feelings to enhance the acceptability of beliefs, needs to be refined for our purposes. At least in the case of intuitive *ethical* judgment, it is the *intensity* of the linked emotional response—regardless of whether it is positive or negative in character—that appears to play a decisive role. The empirical evidence cited above indicates that we are typically most committed to those intuitive judgments that are linked to the strongest emotion; but this research does not provide grounds for attributing this power only to positive feelings. On the contrary, it seems plausible that negative emotions are even more important in this regard. This would be a good question for further empirical research, but one that we can temporarily leave to the side.

Luckily, it is relatively easy to adapt Thagard's HOTCO II model in the way needed. Let us call this revised version HOTCOWRE. Much like HOTCO II, each belief

in the HOTCOWRE model has a valence, in addition to an activation level (which simulates the acceptability of the belief). The main difference is that valences in HOTCOWRE represent the relative intensity of the linked emotion, as opposed to HOTCO II in which valences simulate positive or negative feelings. HOTCOWRE does not distinguish between positive and negative feelings in calculating emotional coherence. With this model, each belief in the network has a valence between 0 and 1, with 0 representing no emotional intensity and 1 representing the highest degree of emotional intensity. Beliefs are initially admitted into normative consideration with a valence that corresponds to the level of emotional conviction with which they are held prior to philosophical reflection; and these valences determine the initial activation level of the beliefs to which they are linked. Otherwise, valences and activation levels in HOTCOWRE are calculated in the same manner as in HOTCO II. Initial valences and activation levels are adjusted based on emotional and explanatory coherence considerations. The activation level and valence of a particular unit is a function of the activation levels and valences of all the units to which it is connected; and the activation level of a unit is determined in part by its valence. HOTCOWRE nicely simulates the impact of emotional intensity on the acceptability of beliefs, in line with the empirical evidence outlined in this project. According to this model, beliefs that are linked to stronger feelings (i.e., have a higher valence) are more likely to be endorsed in wide reflective equilibrium, as emotional coherence considerations are factored into the overall coherence calculations that determine normative acceptability.

As noted above, the coherentist method endorsed here affords moral emotion justificatory force in a more modest way than Classical Intuitionism. Like Rawls and

Daniels, Classical Intutionists, such as Pritchard and Ross, failed to adequately account for the important connection between moral intuition and emotion. Again, it appears that the strong emotion characteristically linked to our moral intuitions is what causes us to be so deeply committed to them. Classical Intuitionists took the legitimacy of these pre-reflective judgments for granted, asserting that our moral intuitions are, in fact, morally justified, just as we tend to believe. By affording our moral intuitions foundational normative weight, *Classical Intuitionists unwittingly afforded a similar degree of justificatory force to the feelings that underlie these convictions.* Although they were unaware, Pritchard and Ross attributed a foundational normative weight to our gut feelings by automatically endorsing the beliefs to which they are connected. By contrast, the method of wide reflective affords justificatory force to our moral feelings in a more balanced way. While the feelings linked to a moral belief carry justificatory force (by directly impacting the normative weight of the associated belief), neither these feelings nor the intuitions to which they are linked are automatically deemed justified. Beliefs with stronger emotional backing are more likely to be endorsed in wide reflective equilibrium, but this is no guarantee. The method of wide reflective equilibrium balances 'hot' emotional considerations with 'cold' explanatory ones in the overall calculation that determines normative acceptability. It is likely that some beliefs that have a high valence when they are first admitted into normative consideration will fail to be among the set ultimately endorsed in wide reflective equilibrium. The fact that this is a genuine likelihood marks a significant departure from Classical Intuitionism, a position which automatically justifies our moral intuitions and feelings.

This section has highlighted one of the two main senses in which the method of wide reflective equilibrium is a naturalistic approach to moral justification by underscoring how it affords our core moral intuitions and feelings justificatory force. HOTCOWRE was endorsed as a psychologically realistic model that incorporates emotional coherence considerations as part of the overall calculation that determines belief acceptability. According to this model, the normative weight of beliefs is determined in part by the intensity of the emotion to which they are connected. This added feature of the method of wide reflective equilibrium--which both Rawls and Daniels failed to incorporate—fits with empirical research demonstrating a strong link between moral feeling and conviction. *By incorporating emotional coherence, this coherentist approach ensures that the set of beliefs endorsed in wide reflective equilibrium will largely accord with our gut feelings, without automatically justifying all of them.* To my awareness, there has not been another *developed* theory in the ethical tradition outlining how our gut feelings should be attributed justificatory force as part of a normative reasoning procedure—although this appears to be a necessary requirement for a psychologically realistic approach to moral justification. The next section, in turn, focuses on the other main naturalistic feature of this coherentist method, which also serves to distinguish it from mainstream approaches in the tradition.

III. The Justificatory Force of Scientific Considerations

The method of wide reflective equilibrium affords scientific theories normative weight, as one of two (along with metaethical theories) main types of level (c) considerations. This is an especially distinctive feature for an approach to ethical

justification, since the traditional view in moral philosophy is that empirical considerations are without justificatory force. Proponents of this view typically espouse what has come to be known as the 'is/ought' distinction, citing arguments articulated by David Hume (1964), G.E. Moore (2010) and others. In his *Treatise,* Hume contended that moral 'oughts' cannot be *formally deduced* solely from 'matters of fact' (pg. 469). Years later, Moore argued that moral 'goodness' is a "non-natural quality," which is irreducible to any "naturalistic property," such as 'pleasantness' or 'desirability.' Accordingly, he insisted that any attempt to define 'goodness' as one of these natural properties commits the "naturalistic fallacy." In more contemporary writings, proponents of the 'is/ought' distinction often cite the dictum, 'what is natural is not necessarily right.' In general, I am sympathetic to the idea that moral principles cannot be *formally deduced* (as defined in logic) solely from descriptive premises. Nonetheless, I reject many of the broader implications that proponents of the 'is/ought' distinction have drawn from this basic insight: for example, the notion that empirical considerations and facts are normatively irrelevant.

Indeed, this idea conflicts with an ideal internal to the moral philosophical tradition—the 'ought implies can' principle. Most contemporary philosophers endorse this basic ideal, which has been interpreted in various ways. The basic notion, however, is that moral requirements must be realistic, since it makes little sense to prescribe ethical behaviors or ideals that are impossible to achieve. This ideal implies that there is an important connection between descriptive facts and normative values. Although the relationship is not a *deductive* one, facts about what is possible for human beings *constrain* the range of viable normative principles. According to the 'ought implies can'

principle, unrealistic normative principles should be rejected, since they are of little practical value. This general ideal, however, is underspecified. Philosophers can concur that our normative theories need to be realistic, while disagreeing about the implications of this commitment. Presumably everyone would agree, for example, that moral principles cannot prescribe actions that violate the laws of physics, but this is not a very robust standard. A more compelling idea is that our normative theories also need to be *psychologically* realistic; an insight captured by Owen Flanagan's "Principle of Minimal Psychological Realism" (PMPR), which stipulates, "make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for creatures like us" (Flanagan, 1991, p. 32). Flanagan contends that this principle has been widely accepted (at least tacitly) by moral philosophers regardless of their philosophical orientation. He writes, "almost all traditions of ethical thought are committed to [this] minimal sort of psychological realism" (pg. 32).

For our purposes, the central point is that 'ought implies can' principles, such as Flangan's PMPR, afford at least some types of empirical considerations (e.g., those pertaining to our moral psychology) normative weight. According to this general ideal, relevant scientific considerations constrain the range of viable normative principles, since we must reject prescriptions that are unrealistic. This basic principle is built into the method of wide reflective equilibrium. By incorporating scientific considerations as level (c) considerations, this coherentist approach allows relevant empirical theories to constrain the range of acceptable (a) judgments and (b) principles. Since the justification of (a) and (b) beliefs is based on their overall coherence with considerations at all three

levels, including (c) scientific theories, it would be highly unlikely for empirically implausible (a) and (b) beliefs to be among the set endorsed in reflective equilibrium. This coherentist approach, however, goes further than what is minimally required by the general 'ought implies can' principle, an ideal which might only afford empirical considerations *negative* justificatory force (i.e., as providing grounds only for the rejection, as opposed to endorsement, of prescriptive views). According to the method of wide reflective equilibrium, if a scientific theory coheres better with some (a) judgments and (b) principles as opposed to others, this provides a reason to endorse the more empirically plausible set. In other words, this coherentist method also affords empirical considerations *positive* justificatory force. It would be beyond our scope to offer a detailed argument in favor of this naturalistic feature. Instead, my primary goal in this section is to provide some specific examples of how empirical theories can play an important justificatory role in this coherentist method, in response to Paul Thagard's misguided criticism (2010) that this approach lacks scientific grounding.

Thagard criticizes "narrow" (see Section I above regarding the distinction between "narrow" and wide") versions of the method of reflective equilibrium, while endorsing a Neo-Aristotelian, 'needs-based' approach to moral justification, another method that affords scientific considerations normative weight. Several contemporary, empirically-minded philosophers, including Larry Arnhart (1998) and William Casebeer (2005), have proposed variants of this approach, which is inspired by Aristotle's proposal that human beings have natural ends and that moral 'goodness' or 'rightness' may be defined on this basis. In comparison to Aristotle's original theory—which focused more on the cultivation of virtuous traits, as opposed to the normative justification of discrete

moral judgments and principles--the Neo-Aristotelian views proposed by empirically-minded theorists are typically more consequentialist in orientation. Proponents of this Neo-Aristotelian approach argue that determinations of moral justification should be based on an empirical account of core human needs. Moral principles and practices that are conducive to the satisfaction of our basic needs are 'right' or 'justified.'

Thagard (2010) offers a moral theory exemplifying this consequentialist-style, Neo-Aristotelian approach. Based on a review of relevant empirical research, he argues that, in addition to our basic biological requirements for food, health and security, human beings have other "vital needs." He writes, "love, work and play are not arbitrary wants, but are closely tied to vital human needs for relatedness, competence and autonomy" (pg. 166). Thagard proceeds to evaluate ethical principles on the basis of how well they satisfy these basic needs. For example, he argues that the Classical Utilitarian principle that 'one should not privilege the welfare of friends and loved ones over strangers' violates our fundamental need for relatedness, and hence this moral principle must be rejected. He writes, "if people have a deeper need for relatedness based on close family relationships than for relatedness based on more casual acquaintances… we cannot expect them to put aside a special concern for the well-being of their loved ones" (pg. 200). Thagard summarizes his Neo-Aristotelian approach as follows, "we can assess actions as right or wrong according to how well they satisfy human needs, especially vital needs such as material subsistence, but also social and psychological needs such as relatedness an autonomy…an action is right to the extent that it furthers those needs, and wrong to the extent that it damages them" (pg. 200-207).

It will be argued below that, as a foundationalist approach, this Neo-Aristotelian method suffers from the same basic weakness as other approaches of this type—i.e., an inability to satisfactorily establish the legitimacy of its supposedly 'self-evident' normative foundations and related problems of underdetermination. Before turning to this issue, however, we will focus on Thagard's (2010) criticism of the method of *narrow* reflective equilibrium; which will help to clarify how the method of *wide* reflective equilibrium affords scientific considerations substantial justificatory force. In his 2010 book, Thagard characterizes Rawls' approach as follows, "the method consists of reflectively adjusting our moral intuitions and moral principles until equilibrium is reached in the form of a rich set of intuitions and principles that fit well with each other" (pg. 202). It is clear from this quote that Thagard only has narrow, two-level versions of Rawls' method in mind here. Indeed, in so far as he fails to specify that there are more complex versions of Rawls' method that overcome the weaknesses of narrower varieties—an oversight which is especially ironic given his previous endorsement (2002) of the method of *wide* reflective equilibrium (see Section II above)--Thagard can justly be accused of constructing a straw man. Nonetheless, his criticism of narrow reflective equilibrium is instructive. He articulates a common complaint (see, for example, Hare and Brandt) that Rawls' coherentist method relies too heavily on moral intuitions, which are normatively suspect. He writes,

> The…problem [with Rawls' method] is the highly subjective nature of moral intuitions…We have little idea why we have the particular emotional responses that we do to different situations… Many contemporary ethicists like to treat moral intuitions as evidence, akin to experimental data…[but] moral intuitions have no similar robustness and therefore should not be treated as data. There is thus no reason why they should be allowed as input to the process of reflective equilibrium, even if the consideration of principles can be expected to lead to the revisions of intuitions (pg. 202).

In accord with this argument, Thagard emphasizes that merely reaching an equilibrium state in which (unreliable) intuitions are balanced with consistent principles provides scant grounds for justification. He writes, "the method of reflective equilibrium is flawed because it is often much too easy to reach equilibrium…[People] can settle into equilibrium states with a good fit of intuitions and principles that nevertheless are not very logical" (pg. 202). Thagard indicates that narrow versions of the method of reflective equilibrium are merely sophisticated examples of Classical Intuitionism, which, unlike the Neo-Aristotelian approach he endorses, fail to adequately incorporate scientific considerations. He underscores, "we need a…way to break out of the circle of intuitions and principles that the method reflective equilibrium generates. Vital needs provide the most attractive direction, because the question of what we need to function minimally and maximally as human beings is at least partly empirical" (pg. 203). Accordingly, he emphasizes that identifying vital needs is primarily an empirical pursuit, involving a variety of sciences. "For a broader account of successful functioning as a human being,

we need to look to other empirical sources such as psychology, anthropology, and sociology" (pg. 203).

As noted above, Thagard's criticism of reflective equilibrium only addresses narrow versions of this approach. Indeed, Daniels draws a sharp distinction between "narrow" and "wide" reflective equilibrium, while stressing the inadequacy of the former. He underscores that, in comparison to narrow versions, which only incorporate two types of normative considerations, (a) judgments and (b) principles, wide reflective equilibrium adds an additional level (c), consisting of metaethical and empirical background theories. Daniels writes,

> [a] two-tiered view of moral theories has helped make the problem of theory acceptance or justification in ethics intractable…To be sure, appeal to elementary coherence (here, consistency) between principles and judgments sometimes allows us to clarify our moral views or to make progress in moral argument. But there must be more to justification of both judgments and principles than such simple coherence considerations …It is because *narrow* reflective equilibrium allows no further opportunities for revision that it is readily assimilated to the model of a sophisticated [version of Classical] Intuitionism (1979, pg. 257).

Hence, Daniels shares Thagard's worry about the limitations of narrow reflective equilibrium, i.e., that this method lacks the necessary resources to substantially challenge and revise our intuitive beliefs. Daniels argues, however, that *wide* reflective equilibrium overcomes this weakness by incorporating metaethical and scientific considerations into the justificatory process. He underscores, "*wide* reflective equilibrium…allows for far more drastic *theory-based* revisions of [a] moral judgments…[They] are always subjected

to exhaustive review and are 'tested,' as are the [b] moral principles, against a relevant body of [c] theory" (pg. 266-267). In general, according to the method of wide reflective equilibrium, level (a) judgments and (b) principles are justified not only on the basis of their overall coherence with each other—which would constitute only a narrow reflective equilibrium--but with level (c) considerations as well. In this sense, these level (c) theories and beliefs serve as an 'outside' source of justification, providing a powerful 'check and balance' for the acceptance of (a) and (b) intuitions. Hence, Thagard's criticism of *narrow* reflective equilibrium misses the mark against the wider version endorsed here.

Before delving further into this issue, however, I want to briefly consider another question pertaining to level (c) considerations in the method of wide reflective equilibrium. The discussion so far has focused only on their role in establishing level (a) judgments and (b) principles. Daniels indicates, however, that level (c) considerations may be justified through this coherentist reasoning procedure as well, although he does not develop the idea. For example, he writes, "suppose that some set of considered moral judgments plays a role in constraining the background theories in (c). It is important to note that the acceptability of (c) may thus in part depend on some *moral* judgments…" (1979, pg. 260). Allow me to clarify Daniels' claim that (c) considerations, and not just (a) and (b) ones, can be justified through the method of wide reflective equilibrium. It seems that of the two types of (c) theories, metaethical and scientific, only the former is subject to justification via this coherentist reasoning procedure. Clearly the method of wide reflective equilibrium cannot establish the legitimacy of scientific theories—i.e., *as being good scientific theories*. There are standards internal to science for making this

determination. With regard to metaethical theories, however, Daniels appears to be on firmer ground. Based on this coherentist approach, metaethical theories are justified when they cohere with the set of moral judgments, principles and background theories, including the empirical ones, endorsed in wide reflective equilibrium. There has been much debate in the metaethical literature concerning the proper way to justify metaethical theories. Although it would be beyond our scope to argue the point here, this seems like a very powerful approach. In spite of all the disagreement, presumably every moral philosopher would concur that our metaethical theories need to at least cohere with the ethical judgments and principles we endorse—an outcome which is guaranteed by this coherentist method.

Returning back to the issue above, let us consider some examples of how empirical considerations may play an important justificatory role in the method of wide reflective equilibrium. Daniels suggests one possible avenue, emphasizing that, if the justification for some moral judgments and principles relies on a particular metaethical conception, arguments threatening this level (c) conception may also undermine the related (a) and (b) beliefs. He writes, "[another] possible benefit of wide reflective equilibrium is that level [c] disagreements about theories may be more tractable than disagreements about moral judgments and principles. Consequently, if the moral disagreements can be traced to disagreements about [c] theory, greater moral agreement may result" (1979, pg. 263). Daniel Little (1984) is critical of Daniels contention that settling metaethical disputes may help to resolve disagreements at levels (a) and (b) in wide reflective equilibrium, because philosophers are no more likely to agree about metaethical issues. Little argues, "it seems as reasonable to suppose that broadening the

discussion from moral judgments and principles to moral judgments, principles, and philosophical theories, has simply broadened the possible sources of irresolvable disagreement" (pg. 381). In making this argument, Little fails to recognize that scientific considerations have important implications for at least some types of metaethical issues, like the ones addressed in Chapters 3 and 4, which helps to make *these* metaethical disputes more tractable. This bolsters Daniels' claim that level (c) metaethical considerations may help to adjudicate competing beliefs at levels (a) and (b). Indeed, it appears that level (c) scientific considerations with metaethical implications may carry substantial normative weight in the method of wide reflective equilibrium.

A nice example is provided by Joshua Greene and Jonathon Cohen (2004). They contend that neuroscientific findings indicate that we lack free will, and this (metaethical) finding challenges our intuitive sense of retributive justice. While I agree with many of the key assumptions of their argument, it would be beyond our scope to try to establish these central premises. Instead, what is of primary interest to us is the basic structure of Greene and Cohen's argument, as this relates to the method of wide reflective equilibrium. Greene and Cohen provide the following summary of their position, "free will as we ordinarily understand it is an illusion generated by our cognitive architecture. Retributivist notions of criminal responsibility ultimately depend on this illusion, and if we are lucky, they will give way to consequentialist ones, thus radically transforming our approach to criminal justice" (pg. 1784). As evident from this passage, Greene and Cohen focus specifically on criminal responsibility and punishment, but they evaluate these practices from a moral perspective. They suggest that a retributivist, 'eye for an eye' approach to blame and punishment is not ethically justified, since this approach

relies on an unrealistic metaethical conception of the person—i.e., a view that we posses free will, in the libertarian sense of the term. According to the libertarian account, free will requires the capacity to do or decide otherwise at the moment of choice. In contrast, compatibilist views of free will (see, for example, Dennett, 2003) deny that we have a capacity to do otherwise, while offering a different standard for freedom of action (e.g., acting in accord with our beliefs and desires, as reason-responsive individuals). As noted previously, in addition to intuitive (a) judgments and (b) principles, we also have metaethical intuitions. Greene and Cohen claim that one such metaethical intuition is our instinctual belief that we possess libertarian free will. They argue, however, that neuroscientific findings and other scientific considerations indicate that all of our actions and decisions are fully determined by antecedent causes or brain states, and as such libertarian free will is merely an illusion (for a similar argument, see Pinker, 1997). They write, "the combined effects of genes and environment determine all of our actions…[a finding that] really does threaten free will and responsibility as we intuitively understand them" (pg. 1780).

Greene and Cohen contend that this scientific denial of libertarian free will undermines the moral justification for retributivist approaches to blame and punishment, an approach which holds intuitive appeal for many of us. They write, "retributivism captures the intuitive idea that we legitimately punish to give people what they *deserve* based on their past actions—in proportion to their 'internal wickedness,' to use Kant's phrase…and not, primarily, to promote social welfare in the future" (pg. 1776). Indeed, Greene and Cohen contrast this "backward-looking" approach to moral responsibility and punishment with "forward-looking," consequentialist-utilitarian views, which hold that

"punishment is justified by its future beneficial effects" (pg. 1776). Retributivist punishment is supposed to be a justified form of revenge—a means of 'making things right' or 'meting out just desserts'--in which the guilty are made to pay or make amends for their misdeeds. As suggested by Greene and Cohen, our retributivist tendencies appear to run deep. In support of this view, Frans de Waal (2006) has documented reciprocal, tit-for-tat behaviors, including those which are revenge-like, in apes. For example, chimpanzees are more likely to share food with individuals who have previously groomed them, and are more likely to join in conflicts against individuals who have intervened against them in the past—which may be a form of proto-revenge. Regardless of their origins, our retributivist tendencies have been well-documented throughout history (consider, for example, the Hatfields and the McCoys, or the contemporary Israel-Palestine conflict), and have been expressed in the form of ethical judgments and principles. For example, a victim or witness to a violent crime may issue an (a) judgment that the perpetrator 'deserves to be punished, without mercy.' We also espouse a variety of 'eye for an eye' (b) moral principles, such as 'one wrong turn deserves another,' 'if you stab me in the back, then I can stab you in the back,' 'those who are guilty should suffer to the same extent as their victims,' etc.

Greene and Cohen contend that the justification for these retributivist judgments and principles relies on an unrealistic, libertarian conception of personal accountability and guilt. We believe that people deserve to be punished, because we think they are free, in a libertarian sense of the term. Once it is recognized, however, that the decisions people make are causally determined, this notion of 'just dessert' loses force. Greene and Cohen write, "intuitively, we want to punish those people who truly deserve it, but

whenever the causes of someone's bad behavior are made sufficiently vivid, we no longer see that person as truly deserving punishment" (pg. 1783). According to these authors, we can longer blame the wicked for their misdeeds, at least in the way required to justify retributivist punishment, since it appears that their wickedness is caused, at least in large part, by forces outside of their control—genetic endowment, early environment, etc. Dennett (2003) refers to this as the "creeping specter of exculpation" that follows from a scientific understanding of human behavior. Greene and Cohen argue, however, that we can still justifiably hold people accountable and punish them for their misdeeds within a deterministic framework, but only on consequentialist-utilitarian grounds (i.e., for the purpose of behavior modification, deterrence, and the protection of others). They conclude, "retributivism…ultimately depends on an intuitive, libertarian notion of free will that is undermined by science. Therefore, with the rejection of common-sense conceptions of free will comes the rejection of retributivism and an ensuing shift towards a consequentialist approach to punishment, i.e., one aimed at promoting future welfare rather than meting out just desserts" (pg. 1776).

Regardless of whether their argument is a good one or not, Greene and Cohen have provided a prime example of how scientific considerations may play an important justificatory role in the method of wide reflective equilibrium. The authors contend that the justification for retributivist (a) judgments and (b) principles relies on a (c) metaethical conception of the person (i.e., that we possess libertarian free will), which is undermined by scientific theory and evidence. They argue on this basis that we ought to reject these groundless intuitive biases, and instead endorse a set of consequentialist-utilitarian judgments and principles that are more consistent with a deterministic view of

the person. Greene and Cohen argue specifically that consequentialist, forward-looking justifications for punishment are the only ones that make sense within this deterministic framework. Based on this example, it is clear that scientific findings with metaethical implications may impact the normative weight of level (a) and (b) judgments whose justification relies, at least in part, on these underlying metaethical conceptions. Let us call this the "metaethical route" for affording scientific considerations normative weight in this coherentist method.

Empirical theories may also have a more direct impact (i.e., without reference to underlying metaethical conceptions) on the justification of moral (a) judgments and (b) principles. Consider, for example, the implications of a Darwinian genealogy of our moral sensibility. According to this evolutionary account, many of our core intuitions are expressions of evolved affective biases that were fitness-enhancing for our ancestors, and this is why they are part of our cognitive make-up today. As emphasized in Chapter 4, this scientific theory poses a serious epistemological challenge to moral realism, a thesis that there are 'independent' moral truths or facts that apply irrespective of our subjective desires. Of course, ethical intuitions and beliefs may be justified, even if they are not objectively true. The method of wide reflective equilibrium, for example, is an approach to moral justification that makes no appeal to objective truths. Nonetheless, it would seem that this Darwinian account also provides grounds for *questioning* the justification of these intuitions. This evolutionary genealogy implies that we are confident about these beliefs *because this was fitness-enhancing*, which does not require that they are ethically justified. Although this theory does not necessarily imply these evolved beliefs are *unjustified*¸ at the very least, it raises doubts. Indeed, there is an interesting tension

here. On the one hand, as emphasized in Section II, the method of wide reflective equilibrium affords our moral feelings normative weight by incorporating emotional coherence considerations as part of the overall calculation that determines belief acceptability, such that intuitions with greater emotional backing are more likely to be endorsed. On the other hand, this coherentist method also affords relevant scientific theories normative weight, and one of those theories explains why we are naturally committed to our gut intuitions without any reference to their truth or justification.

This internal tension, however, is neither a problem for this justificatory method nor most of the evolved intuitions it incorporates. The method of wide reflective equilibrium requires only that our moral intuitions and the associated gut feelings be afforded normative weight, and not that they must ultimately be deemed justified. In principle, this method could lead to the rejection of most of our evolved intuitions, but this would be an almost inconceivable result. As outlined in Section I, there will be 'provisional fixed points' in this justificatory network, many of which will be evolved intuitions. These 'core' beliefs, such as 'murder is wrong' or 'helping others is good,' have strong emotional backing and are so deeply intertwined with the moral judgments, principles and metaethical conceptions of which we are characteristically most confident as to be virtually untouchable. Again, according to this coherentist method, the justification of a belief is determined by its overall *explanatory and emotional coherence* with considerations at all three levels. Hence, although a Darwinian genealogy of our moral sensibility raises doubts about the justification of our evolved beliefs and the legitimacy of our deep emotional commitment to them—and thus would diminish, at least to some degree, the normative weight of these beliefs—this scientific theory will

not, by itself, suffice for deeming them unjustified. As noted above, this evolutionary account does not necessarily imply that our core intuitions are unjustified, and there will be many other coherence considerations weighing in their favor. While it is plausible that some of our evolved beliefs—for instance, those pertaining to retributivism—may ultimately be rejected in wide reflective equilibrium, this would be based on considerations in addition to the doubts raised by a Darwinian genealogy. In general, for the reasons cited above, it seems highly unlikely that *most* of our core intuitions would be deemed unjustified at the end of this coherentist reasoning process. Nonetheless, even though these (c) evolutionary considerations are unlikely to play a decisive role, they have a direct impact on the normative weight of our evolved intuitions. This demonstrates another way (in addition to the "metaethical route") that scientific considerations my carry justificatory force in the method of wide reflective equilibrium: by directly lowering (or enhancing) the justificatory force of level (a) and (b) beliefs. Let us call this the "direct relevance route."

Another example of the direct relevance route is provided by Thagard (2010). The method of wide reflective equilibrium would also readily incorporate the types of empirical findings (i.e., those concerning our vital needs) that Thagard utilizes in his Neo-Aristotelian approach to moral justification. Empirical research regarding our basic needs can directly bear on the normative weight of relevant (a) judgments and (b) principles. For example, a study indicating that a federal ban on active euthanasia causes significant psychological harm to terminally ill patients and their loved ones would bolster moral judgments and principles that challenge this ban. Whether or not these ethical beliefs in favor of active euthanasia were ultimately endorsed, however, would

depend on how well they cohere with the set of beliefs and principles endorsed in wide reflective equilibrium. There could be other considerations (e.g., a moral belief that "we ought to preserve human life whenever possible") weighing against the endorsement of euthanasia practices.

Indeed, this is one of the chief strengths of the method of wide reflective equilibrium in comparison to Neo-Aristotelian approaches of the sort endorsed by Thagard. The coherentist reasoning procedure defended here is designed to balance competing considerations and make subtle discriminations; whereas, characteristic of foundationalist approaches to moral justification, Neo-Aristotelianism is underspecified. In order to avoid making highly controversial claims about 'human nature' and our 'proper functioning', these theories, such as Thagard's, typically offer very broad characterizations of our vital needs. As a result, these basic needs, which are supposed to play a foundational role by allowing us to justify a full range of ethical judgments and principles, are too limited for the task. Consider, for example, the basic 'rights' identified by Thagard, such as a need for autonomy (work), relatedness (love) and competence (play). It can be quite difficult to apply these very general principles to concrete cases. These principles can conflict (e.g., work vs. family obligations) and it is not always clear what vital need is most pertinent to a moral issue (e.g., is active euthanasia primarily an autonomy or relatedness issue?). Indeed, Thagard acknowledges this challenge, writing,

> Unfortunately, recognition of human [needs] does not provide an easy answer concerning what to do in cases where it may be necessary to violations of one person's rights in order to prevent the violation of the rights of others….The difficulty of arriving at indisputable moral principles is the result…of the huge complexity of determining the range and importance of human psychological needs and calculating the consequences of the available range of actions" (pg. 199-203).

Hence, it seems that any adequate approach to moral justification must do more than simply identify basic needs. At the very least, it must also outline a reasoning procedure for determining their proper ranking (i.e., by ethical importance) and adjudicating conflicts. The method of wide reflective equilibrium is much better suited to this task than the simple foundationalist reasoning procedure (i.e., deduce principles from basic needs, and then apply these principles to specific cases) outlined by Thagard. Although normatively relevant, empirical findings regarding our basic needs will not suffice for addressing these more nuanced moral questions, which requires a more comprehensive approach, like the method endorsed here.

Moreover, the Neo-Aristotelian approach endorsed by Thagard also rests on questionable normative foundations, which is another typical weakness of foundationalist views. Neo-Aristotelianism relies on a supposedly 'self-evident' metaethical conception concerning what morality is fundamentally about—i.e., protecting and satisfying our basic needs—and defines moral 'rightness' or 'wrongness' on this basis. Thagard suggests that once we can identify our vital needs with the help of empirical science, this can provide a foundation for justifying a full set of moral judgments and principles; but

what justifies the metaethical conception of morality underwriting this approach? Thagard merely appeals to its self-evidence. Over the years, however, philosophers have proposed many different conceptions of what morality is fundamentally for or about (e.g., living in accord with reason, promoting social cohesion, extinguishing desire, achieving self-actualization, etc.), some more appealing than others. At the very least, then, we need an argument for why this Neo-Aristotelian view is more justified than the alternatives, since scientific considerations alone will not suffice for solving *this* metaethical problem. Empirical research can help us to determine what our fundamental needs are, but not that morality primarily concerns this question. Presumably, proponents of this Neo-Aristotelian metaethical conception would point to the fact that it accords with many of our intuitive moral judgments and principles (i.e., many of them appear to relate to the protection and furthering of our basic needs, broadly construed). This type of argument, however, would be an example of the method of wide reflective equilibrium in action, an argument which would avoid the sort of foundationalist claim to metaethical justification hampering Thagard's method. For what it is worth, the conception of morality underwriting the Neo-Aristotelian approach seems attractive, in light of the coherence considerations mentioned above. I remain agnostic, however, about whether or not this is the (c) theory that would emerge in wide reflective equilibrium. Indeed, there appears to be no reason why this coherentist method should result in the endorsement of just one of these metaethical conceptions of morality, since some of these views may complement each other. For example, the Neo-Aristotelian theory may cohere with the idea that morality's primary function is to promote social cohesion, since

it could be argued that a cohesive society is one in which the basic needs of most of its members are satisfied.

As emphasized throughout this chapter, the method of wide reflective equilibrium differs substantially from traditional, foundationalist approaches to moral justification. Foundationalist methods, such as the Neo-Aristotelian one outlined above, typically identify a set of normative considerations that are supposed to be self-evidently true and robust enough to justify a full set of moral beliefs and principles. The main weakness of this general approach is its characteristic reliance on an outmoded, moral realist conception of ethical truth, which was challenged in Chapter Four. For this reason, these foundationalist views fail to adequately justify their normative foundations, which are also typically underspecified. The method of wide reflective equilibrium avoids these difficulties. According to this approach, moral beliefs are justified based on their overall coherence within a network of normative considerations, none of which are automatically endorsed prior to this justificatory procedure. We begin by collecting those beliefs of which we are most confident and then evaluate their emotional and explanatory coherence. In contrast to Classical Intuitionism, the method of wide reflective equilibrium does not take the justification of our moral intuitions and feelings for granted. Rather, the normative weight of beliefs in the network is subject to change based on coherence considerations. Beliefs that are initially admitted with greater emotional backing are more likely to be endorsed in wide reflective equilibrium, but this is not guaranteed; as this coherentist method has the necessary resources to challenge our pre-reflective convictions.

The method of wide reflective equilibrium is also distinctively naturalistic in comparison to other justificatory approaches in the tradition. Most moral philosophers deny that scientific considerations are normatively relevant, and none that I am aware of have provided a detailed account of how our gut feelings should be afforded justificatory force as part of a normative reasoning procedure. In addition to attributing normative weight to scientific considerations, the method of wide reflective equilibrium also incorporates emotional coherence in the overall calculation that determines belief acceptability; which, I have argued, is a necessary feature for a psychologically realistic approach to moral justification. The empirical evidence regarding the central role of emotion in moral judgment and motivation indicates that we cannot be expected to accept and act upon a set of beliefs that do not largely with our gut feelings. By attributing normative weight to our emotional commitments, the method of wide reflective equilibrium ensures a psychologically realistic result, while fulfilling Rawls' original promise to normatively "investigate our moral sensibility."

IV. The Future of Normative Ethics

As noted at the outset of this project, David Hume famously wrote that "reason is, and ought only to be the slave of the passions." This passage includes both a descriptive and a normative claim. With regard to the former, it appears that Hume was definitely on the right track. The empirical evidence reviewed here supports his basic contention that emotion plays a pivotal role in moral cognition. Affect drives our intuitive assessments, grounds our empathic capacities, guides our everyday moral reasoning, and invests our ethical judgments with practical clout. Hume's claim that moral reasoning is a "slave" to

the passions is probably too strong, however. It remains an open question to what degree our evolved intuitions and affective biases are cognitively penetrable, but there is currently insufficient evidence to conclude that normative moral reasoning lacks the power to impact and redirect our intuitive commitments. The guiding assumption in this final chapter, based on the available evidence, is that in order to be psychologically realistic (i.e., acceptable and motivating) the set of beliefs endorsed in wide reflective equilibrium must *largely cohere* with our gut feelings. This does not require, however, the endorsement of *all* of our intuitive beliefs. Nor does it entail that normative moral reasoning is incapable of impacting our emotional commitments. Indeed, the justificatory reasoning model sketched above (HOTCOWRE) assumes that our gut feelings can be strengthened or weakened based on 'cold' considerations of explanatory coherence. This seems like a reasonable assumption given the current state of evidence, but clearly the cognitive penetrability of our gut feelings and intuitions is an area deserving of much greater attention and study. Indeed, in my view, this is of one of the most important directions for future research in the field of descriptive ethics. There is compelling evidence that evolved emotion and intuition drive our everyday decision-making, but almost no research regarding what happens when individuals are made aware of these biases and their origins. For example, would debriefing subjects of the Trolley Problem studies regarding the evolutionary explanation for our divergent reactions to the Standard and Footbridge Scenarios (i.e., that only the latter triggers our evolved aversion to close/personal harm) alter their judgment style for similar scenarios in the future? It seems that studies like these could readily be developed, and they are most certainly

needed to test Hume's stronger claim that we are slavishly bound by our emotional biases.

The other facet of Hume's famous quote, that reason *ought* to be the slave of the passions, has received relatively less attention from empirically-minded ethicists, despite its great level of importance for normative ethics. Again, I do not find the 'slave' language very apt, but the thrust of Hume's contention, as I interpret it, is that normative moral theories need to be psychologically realistic; which, in turn, requires that they substantially overlap with our gut feelings. In accord with this insight, my primary aim in this chapter was to outline a psychologically realistic justificatory reasoning procedure. Due to space constraints, however, I was unable to develop an adequately detailed argument for *why* our normative moral theories ought to meet this standard. I endorsed Flanagan's Principle of Minimal Psychological Realism (PMPR)—which stipulates, "make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible…for creatures like us." Flanagan's principle provides grounds for rejecting a sharp 'is/ought' distinction, but it does not clearly specify what a commitment to psychological realism entails. The empirical research regarding moral judgment and motivation reviewed here provides the needed clarity: if the goal is to generate moral prescriptions that are genuinely actionable—i.e., capable of acceptance and behavioral follow-through—then these prescriptions must not run afoul of our core emotional convictions.

This statement of what psychological realism requires does not yet, however, establish that this is a necessary feature for a normative theory to possess. Flanagan suggests that all moral philosophers would presumably endorse PMPR, but this does not

seem to be the case historically, at least when the implications of PMPR are properly spelled out. Most ethicists have denied that our gut feelings are normatively relevant, and in the rationalist tradition many have argued that moral prescriptions do not need to be genuinely actionable, in the sense defined above. According to this rationalist view, moral prescriptions should serve as 'regulative ideals' i.e., ideals that are not practically achievable, but nonetheless beneficial to pursue. I think the coherentist approach endorsed here can countenance regulative ideals as well, but this rationalist challenge highlights the complexity of questions regarding the proper role of psychological realism in normative theorizing. Indeed, this general issue plumbs to the very core of normative ethics, by raising questions about the general purpose of the enterprise. Why should theorists engage in normative theorizing? Is the aim to generate theories for a wider audience in the hopes of improving moral behavior, or some other goal? The case for psychological realism in normative ethics hinges on the answer to these programmatic questions. In short, what is required next is a clear articulation of the purpose and goals of normative ethics for an age in which the 'science of morality' is jumping by leaps and bounds. This important pursuit was largely beyond the scope of this project. Nonetheless, I hope to have made some important strides in the right direction, in line with Hume's original effort to better connect the descriptive and normative dimensions of ethical theorizing. In my view, the central question for normative ethics going forward can no longer be whether 'what is natural is right.' Rather, clarifying the obvious interconnection is the main task for the future.