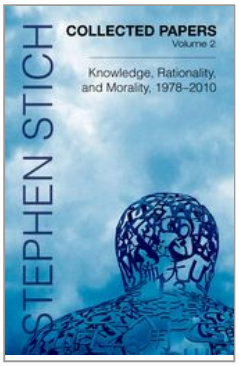

Collected Papers, Volume 2: Knowledge, Rationality, and Morality, 1978-2010
Stephen Stich

Print publication date: 2012
Print ISBN-13: 9780199733477
Published to Oxford Scholarship Online: September 2012
DOI: 10.1093/acprof:oso/9780199733477.001.0001

As a Matter of Fact

Empirical Perspectives on Ethics

John M. Doris
Stephen Stich

DOI:10.1093/acprof:oso/9780199733477.003.0011

Abstract and Keywords

This chapter discusses character and virtue ethics, moral motivation, moral disagreement, and the role of thought experiments in ethics. Though it covers a lot of ground, the leitmotif that runs through the chapter is that there is a growing body of empirical evidence in psychology and neuroscience that philosophers interested in these issues cannot afford to ignore.

Keywords: ethics, moral motivation, moral disagreement, thought experiments

> Too many moral philosophers and commentators on moral philosophy … have been content to invent their psychology or anthropology from scratch….
>
> S. Darwall, A. Gibbard, and P. Railton (1997, 34–35)

1. Introduction

Regarding the assessment of Darwall and colleagues, we couldn't agree more: far too many moral philosophers have been content to *invent* the psychology or anthropology on which their theories depend, advancing or disputing empirical claims with little concern for empirical evidence. We also believe—and we expect Darwall, Gibbard, and Railton would agree—that this empirical complacency has impeded progress in ethical theory and discouraged investigators in the

biological, behavioral, and social sciences from undertaking philosophically informed research on ethical issues.

We realize that some moral philosophers have taken there to be good reasons for shunning empirical inquiry. For much of the twentieth century, many working in analytic ethics—variously inspired by Hume's (1978, 469) pithy injunction against inferring *ought* from *is* and the seductive mysteries of Moore's (1903, esp. 10–17) "Open Question Argument"—maintained that descriptive considerations of the sort adduced in the natural and social sciences cannot constrain ethical reflection without vitiating its prescriptive or normative character (e.g., Stevenson 1944, 108–10; R. M. Hare 1952, 79–93). The plausibility of such claims is both debated and debatable, but it is clear that they have helped engender suspicion regarding "naturalism" in ethics, which we understand, broadly, as the view that *ethical theorizing should be an (in part) a posteriori inquiry richly* **(p.248)** *informed by relevant empirical considerations*.[1] Relatedly, this anti-naturalist suspicion enables disciplinary xenophobia in philosophical ethics, a reluctance to engage research beyond the philosophical literature. The methodology we advocate here—a resolutely naturalistic approach to ethical theory squarely engaging the relevant biological, behavioral, and social sciences—flouts both of these anxieties.

Perhaps those lacking our equanimity suspect that approaches of the sort we endorse fail to heed Stevenson's (1963, 13) advice that "Ethics must not be psychology," and thereby lapse into a noxious "scientism" or "eliminativism." Notoriously, Quine (1969, 75) advocated eliminativism in his rendering of naturalized epistemology, urging philosophical "surrender of the epistemological burden to psychology." Quine was sharply rebuked for slighting the normative character of epistemology (e.g., Kim 1988; Stich 1993a), but we are not suggesting, in a rambunctiously Quinean spirit, "surrender of the *ethical* burden to psychology." And so far as we know, neither is anyone else. Ethics must not—indeed cannot—*be* psychology, but it does not follow that ethics should *ignore* psychology.

The most obvious, and most compelling, motivation for our perspective is simply this: It is not possible to step far into the ethics literature without stubbing one's toe on empirical claims. The thought that moral philosophy can proceed unencumbered by facts seems to us an unlikely one: there are just too many places where answers to important ethical questions require—and have very often presupposed—answers to empirical questions.

A small but growing number of philosophers, ourselves included, have become convinced that answers to these empirical questions should be informed by systematic empirical research.[2] This is not to say that relevant information is easy to come by: the science is not always packaged in forms that are easy on the philosophical digestion. As Darwall et al. (1997, 47 ff.) caution, one won't

often find “a well-developed literature in the social sciences simply awaiting philosophical discovery and exploitation.” Still, we are more optimistic than Darwall and colleagues about the help philosophers can expect from empirical literatures: science has produced much experimental and theoretical work that appears importantly relevant to ongoing debates in ethical theory, and some moral philosophers have lately begun to pursue empirical investigations. To explore the issues fully requires far more space than is available here; we must content ourselves with developing a few rather programmatic examples of how an empirically sensitive philosophical ethics might proceed.

Our point is not that reference to empirical literatures can be expected, by itself, to resolve debates in moral theory. Rather, we hope to convince the reader that these literatures are often deeply relevant to important debates, and it is therefore intellectually **(p.249)** irresponsible to ignore them. Sometimes empirical findings seem to contradict what particular disputing parties assert or presuppose, while in other cases, they appear to reconfigure the philosophical topography, revealing that certain lines of argument must traverse empirically difficult terrain. Often, philosophers who follow these challenging routes will be forced to make additional empirical conjectures, and these conjectures, in their turn, must be subject to empirical scrutiny. The upshot, we conclude, is that an intellectually responsible philosophical ethics is one that continuously engages the relevant empirical literature.

2. Character

In the second half of the twentieth century the “ethics of virtue” became an increasingly popular alternative to the Kantian and utilitarian theories that had for some time dominated normative ethics. In contrast to Kantianism and utilitarianism, which despite marked differences share an emphasis on identifying morally obligatory actions, virtue-centered approaches emphasize the psychological constitution, or character, of actors. The central question for virtue ethics, so the slogan goes, is not what sort of action to do, but what sort of person to be.[3] As Bernard Williams (1985, 1) has eloquently reminded us, the “aims of moral philosophy, and any hopes it may have of being worth serious attention, are bound up with the fate of Socrates’ question” How should one live?, and it has seemed to many philosophers, not least due to Williams’s influence, that any prospects for a satisfying answer rest with the ethics of character. Allegedly, if ethical reflection is to help people understand and improve themselves and their relations to others, it must be reflection focused on the condition and cultivation of character (see Williams 1993, 91–95).

Virtue ethics, especially in the Aristotelian guises that dominate the field, typically presupposes a distinctive account of human psychology. Nussbaum (1999, 170), although she insists that the moniker “virtue ethics” has been used to tag such a variety of projects that it represents a “misleading category,” observes that approaches so titled are concerned with the “settled patterns of

motive, emotion, and reasoning that lead us to call someone a person of a certain sort (courageous, generous, moderate, just, etc.).” If this is a fair characterization—and we think it is—then virtue ethics is marked by a particular interest in moral psychology, an interest in the cognitive, affective, and emotional patterns that are associated with the attribution of character traits.[4] This interest looks to be an empirical interest, and it’s natural to ask how successfully virtue ethics addresses it.

(p.250) The central empirical issue concerns, to borrow Nussbaum’s phrase, “settled patterns” of functioning. According to Aristotle, genuinely virtuous action proceeds from “firm and unchangeable character” rather than from transient motives (1984, 1105a28–b1); while the good person may suffer misfortune that impairs his activities and diminishes happiness, he “will never [*oudepote*] do the acts that are hateful and mean” (1984, 1100b32–34; cf. 1128b29; cf. Cooper 1999, 299 ff.).[5] In an influential contemporary exposition, McDowell (1978, 26–27) argued that considerations favoring vicious behavior are “silenced” in the virtuous person; although such an individual may recognize inducements to vice, she will not count them as reasons for action. As we understand the tradition, virtues are supposed to be robust traits; if a person has a robust trait, she can be confidently (although perhaps not with absolute certainty) expected to display trait-relevant behavior across a wide variety of trait-relevant situations, even where some or all of these situations are not optimally conducive to such behavior (Doris 2002, 18).[6]

Additionally, some philosophers have supposed that character will be evaluatively integrated—traits with associated evaluative valences are expected to co-occur in personality (see Doris 2002, 22; Flanagan 1991, 283–90). As Aristotle (1984, 1144b30–1145a2; cf. Irwin 1988, 66–71) has it, the virtues are inseparable; given the qualities of practical reason sufficient for the possession of one virtue, one can expect to find the qualities of practical reason sufficient for them all.

While understandings of character and personality akin to those just described have been hotly contested in psychology departments at least since the critiques of Vernon (1964), Mischel (1968), and Peterson (1968), moral philosophers have not been especially quick in taking the matter up. Flanagan’s (1991) careful discussion broached the issue in contemporary analytic ethics, while Doris (1998, 2002) and Harman (1999, 2000) have lately pressed the point less temperately: although they manifest some fraternal disagreements, Harman and Doris both insist that the conception of character presupposed by virtue ethics is empirically inadequate.

The evidence for this contention, often united under the theoretical heading of "situationism," has been developed over a period of some seventy years, and includes some of the most striking research in the human sciences.

> • Mathews and Canon (1975, 574–75) found subjects were five times more likely to help an apparently injured man who had dropped some books when **(p.251)** ambient noise was at normal levels than when a power lawnmower was running nearby (80 percent v. 15 percent).
> • Darley and Batson (1973, 105) report that passers-by not in a hurry were six times more likely to help an unfortunate who appeared to be in significant distress than were passers-by in a hurry (63 percent vs. 10 percent).
> • Isen and Levin (1972, 387) discovered that people who had just found a dime were twenty-two times more likely to help a woman who had dropped some papers than those who did not find a dime (88 percent vs. 4 percent).
> • Milgram (1974) found that subjects would repeatedly "punish" a screaming "victim" with realistic (but simulated) electric shocks at the polite request of an experimenter.
> • Haney, Banks, and Zimbardo (1973) describe how college students role-playing in a simulated prison rapidly descended to *Lord of the Flies* barbarism.

There apparently exists an alarming disproportion between situational input and morally disquieting output; it takes surprisingly little to get people behaving in morally undesirable ways. The point is not that circumstances influence behavior, or even that seemingly good people sometimes do lousy things. No need to stop the presses for that. Rather, the telling difficulty is just how insubstantial the situational influences effecting troubling moral failures seem to be; it is not that people fall short of ideals of virtue and fortitude, but that they can be *readily* induced to *radically* fail such ideals.

The argument suggested by this difficulty can be outlined as follows: a large body of research indicates that cognition and behavior are extraordinarily sensitive to the situations in which people are embedded. The implication is that individuals—on the altogether plausible assumption that most people will be found in a range of situations involving widely disparate cognitive and behavioral demands—are typically highly variable in their behavior, relative to the behavioral expectations associated with familiar trait categories such as honesty, compassion, courage, and the like. But if people's behavior were typically structured by robust traits, one would expect quite the opposite: namely, behavior consistent with a given trait—e.g., behavior that is appropriately and reliably honest, compassionate, or courageous—across a diversity of situations. It follows, according to the argument, that behavior is not typically structured by the robust traits that figure centrally in virtue-theoretic

moral psychology. Analogous considerations are supposed to make trouble for notions of evaluative integration; the endemic lack of uniformity in behavior adduced from the empirical literature undermines expectations of integrated character structures.

The situationist argument has sometimes been construed by philosophers as asserting that character traits "do not exist" (Flanagan 1991, 302; Athanassoulis 2000, 219–20; **(p.252)** Kupperman 2001, 250), but this is a misleading formulation of the issue.[7] In so far as to deny the existence of traits is to deny the existence of persisting dispositional differences among persons, the claim that traits do not exist seems unsustainable, and the exercise of refuting such a claim idle. (Indeed, it is a claim that even psychologists with strong situationist sympathies, e.g., Mischel 1968, 8–9, seem at pains to disavow.) The real issue dividing the virtue theorist and the situationist concerns the appropriate characterization of traits, not their existence or nonexistence. The situationist argument that needs to be taken seriously, and which to our mind stands unrefuted, holds that the Aristotelian conception of traits as robust dispositions —the sort which lead to trait-relevant behavior across a wide variety of trait-relevant situations—is radically empirically undersupported. To put the ethical implications of this a bit aggressively, it looks as though attribution of robust traits like virtues may very well be unwarranted in most instances,[8] programs of moral education aimed at inculcating virtues may very well be futile, and modes of ethical reflection focusing moral aspirations on the cultivation of virtue may very well be misguided.

At this point, the virtue theorist may offer one of two responses. She can accept the critics' interpretation of the empirical evidence while denying that her approach makes empirical commitments of the sort the evidence indicates is problematic. Or she can allow that her approach makes commitments in empirical psychology of the sort that would be problematic if the critics' interpretations of the evidence were sustainable, but deny that the critics have interpreted the evidence aright. The first option, we might say, is "empirically modest" (see Doris 2002, 110–12): because such renderings make only minimal claims in empirical psychology, they are insulated from empirical threat. The second option, conversely, is "empirically vulnerable" (see Railton 1995, 92–96): it makes empirical claims with enough substance to invite empirically motivated criticism.

We shall first discuss empirically modest rejoinders to the situationist critique. Numerous defenders of virtue ethics insist that virtue is not expected to be widely instantiated, but is found in only a few extraordinary individuals, and these writers further observe that this minimal empirical commitment is quite compatible with the disturbing, but not exceptionlessly disturbing, behavior in experiments like Milgram's (see Athanassoulis 2000, 217–19; DePaul 1999; Kupperman 2001, 242–43). The critics are bound to concede the point, since the

empirical evidence cannot show that the instantiation of virtue in actual human psychologies is impossible; no empirical evidence could secure so strong a result. But so construed, the aspirations of virtue ethics are not entirely clear; if **(p.253)** virtue is *expected* to be rare, it is not obvious what role virtue theory could have in a (generally applicable) program of moral education.[9] This rings a bit odd, given that moral education—construed as aiming for the development of the good character necessary for a good life—has traditionally been a distinctive emphasis in writing on virtue, from Aristotle (1984, 1099b9–32, 1103b3–26) to Bennett (1993, 11–16; cf. Williams 1985, 10). Of course, the rarity of virtue might be thought a contingent matter; given the appropriate modalities of moral education, the virtue ethicist might say, virtue can be widely inculcated. But philosophers, psychologists, and educators alike have tended to be a bit hazy regarding particulars of the requisite educational processes; theories of moral education, and character education in particular, are typically not supported by large bodies of systematic research adducing behavioral differences corresponding to differing educational modalities (Leming 1997a, b; Hart and Killen 1999, 12; Doris 2002, 121–27).

It is tempting to put the situationist point a bit more sharply. It is true that the evidence does not show that the instantiation of virtue in actual human psychologies is impossible. But it also looks to be the case that the available systematic empirical evidence is compatible with virtue being psychologically impossible (or at least wildly improbable), and this suggests that the impossibility of virtue is an empirical possibility that has to be taken seriously. So while the evidence doesn't refute an empirically modest version of virtue ethics, it is plausibly taken to suggest that the burden of argument has importantly shifted: The advocate of virtue ethics can no longer simply assume that virtue is psychologically possible. If she can't offer compelling evidence—very preferably, more than anecdotal evidence—favoring the claim that virtue is psychologically possible, then she is in the awkward position of forwarding a view that would be undermined if an empirical claim which is not obviously false were to turn out to be true, without offering compelling reason to think that it won't turn out to be true.

Suppose the realization of virtue were acknowledged to be impossible: it might yet be insisted that talk of virtue articulates ethical ideals that are well suited—presumably better suited than alternatives, if virtue ethics is thought to have distinctive advantages—to facilitating ethically desirable conduct (see Blum 1994, 94–96). Asserting such a practical advantage for virtue ethics entails an empirical claim: reflection on the ideals of virtue can help actual people behave better. For example, it might be claimed that talk of virtue is more compelling, or has more motivational "grip," than abstract axiological principles. We know of little systematic evidence favoring such claims, and we are unsure of what sort of experimental designs are fit to secure them, but the only point we need to insist on is that even this empirically modest rendering of virtue ethics may bear

contentious empirical commitments. If virtue ethics is alleged to have practical implications, it cannot avoid empirical assertions regarding the cognitive and motivational equipment with which people navigate their moral world.

(p.254) Even without an answer to such practical questions, it might be thought that virtue ethics is fit to address familiar conceptual problems in philosophical ethics, such as rendering an account of right action. In Hursthouse's (1999, 28; cf. 49–51) account of virtue ethics, "An action is right if it is what a virtuous agent would characteristically (i.e., acting in character) do in the circumstances." Hursthouse (1999, 123–26, 136, 140) further insists that an action does not count as "morally motivated" simply by dint of being the sort of thing a virtuous person does, done for reasons of the sort the virtuous person does it for; it must proceed "from virtue," that is, "from a settled state of good character." If this requirement is juxtaposed with the observation that the relevant states of character are extremely rare, as an empirically modest rendering of virtue ethics maintains, we apparently get the result that "morally motivated" actions are also extremely rare (a virtue-theoretic result, interestingly, with which Kant would have agreed). This need not trouble Hursthouse (1999, 141–60); she seems to allow that very often—perhaps always—one sees only approximations of moral motivation. It does trouble us. We think that less than virtuous people, even smashingly less than virtuous people, sometimes do the right thing for the right reasons, and these actions are fit to be honored as "morally motivated." It may not happen as often as one would like, but morally motivated conduct seems to happen rather more frequently than one chances on perfect virtue. Oskar Schindler, the philandering war profiteer who rescued thousands of Jews from the Nazis, is a famous example of the two notions coming apart (see Kenneally 1982), but with a little attention to the history books, we can surely adduce many more. The burden of proof, it seems to us, is on those asserting that such widely revered actions are not morally motivated.

There are also serious questions about the competitive advantages enjoyed by empirically modest virtue ethics. It has seemed to many that a chief attraction of character-based approaches is the promise of a lifelike moral psychology—a less wooden depiction of moral affect, cognition, motivation, and education than that offered by competing approaches such as Kantianism and utilitarianism (Flanagan 1991, 182; Hursthouse 1999, 119–20). Proponents of virtue ethics, perhaps most prominently MacIntyre (1984) and Williams (1985, 1993), link their approach—as Anscombe (1958, 4–5) did in a paper widely regarded as the call to arms for contemporary virtue ethics—to prospects for more psychological realism and texture. We submit that this is where a large measure of virtue ethics' appeal has lain; if virtue ethicists had tended to describe their psychological project along the lines just imagined, as deploying a moral

psychology only tenuously related to the contours of actual human psychologies, we rather doubt that the view would now be sweeping the field.

We contend that for virtue ethics to retain its competitive advantage in moral psychology it must court empirical danger by making empirical claims with enough substance to be seriously tested by the empirical evidence from psychology. For instance, the virtue theorist may insist that while perfect virtue is rare indeed, robust traits approximating perfect virtue—reliable courage, temperance, and the rest—may be widely **(p.255)** inculcated, and perhaps similarly for robust vices—reliable cowardice, profligacy, and so on.[10] To defend such a position, the virtue theorist must somehow discredit the critic's empirical evidence. Various arguments might be thought to secure such a result: (i) The situationist experiments might be methodologically flawed; problems in experimental design or data analysis, for example, might undermine the results. (ii) The experiments might fail standards of ecological validity; the experimental contexts might be so distant from natural contexts as to preclude generalizations to the "real world." (iii) General conclusions from the experiments might be prohibited by limited samples; in particular, there appears to be a dearth of longitudinal behavioral studies that would help assess the role of character traits "over the long haul." (iv) The experiments may be conceptually irrelevant; for example, the conceptions of particular traits operationalized in the empirical work may not correspond to the related conceptions figuring in virtue ethics.

The thing to notice straight away is that motivating contentions like the four above require evaluating a great deal of psychological research; making a charge stick to one experiment or two, when there are hundreds, if not thousands, of relevant studies, is unlikely to effect a satisfying resolution of the controversy. The onus, of course, falls on both sides: just as undermining arguments directed at single experiments are of limited comfort to the virtue theorist, demonstrating the philosophical relevance of a lone study is not enough to make the critics' day. Newspaper science reporting notwithstanding, in science there is seldom, or never, a single decisive experiment or, for that matter, a decisive experimental failure. General conclusions about social science can legitimately be drawn only from encountering, in full detail, a body of research, and adducing patterns or trends. Doris (2002) has recently attempted to approximate this methodological standard in a book-length study, and he there concludes that major trends in empirical work support conclusions in the neighborhood indicated by the more programmatic treatments of Doris (1998) and Harman (1999, 2000). Whether or not one is drawn to this conclusion, we think it clear that the most profitable discussion of the empirical literature will proceed with detailed discussion of the relevant empirical work. If an empirically vulnerable virtue ethics is to be shown empirically defensible, defenders must provide much fuller consideration of the psychology. To our knowledge, extant defenses of virtue ethics in the face of empirical attack do not approximate the

required breadth and depth.[11] Hopefully, future discussions will rectify this situation, to the edification of defenders and critics alike.

3. Moral Motivation

(p.256) Suppose a person believes that she ought to do something: donate blood to the Red Cross, say, or send a significant contribution to an international relief agency. Does it follow that she will be moved actually to act on this belief? Ethical theorists use internalism to mark an important cluster of answers to this question, answers maintaining that the motivation to act on a moral judgment is a necessary or intrinsic concomitant of the judgment itself, or that the relevant motivation is inevitably generated by the very same mental faculty that produces the judgment.[12] One familiar version of internalism is broadly Kantian, emphasizing the role of rationality in ethics. As Deigh (1999, 289) characterizes the position, "reason is both the pilot and the engine of moral agency. It not only guides one toward actions in conformity with one's duty, but it also produces the desire to do one's duty and can invest that desire with enough strength to overrule conflicting impulses of appetite and passion." A notorious difficulty for internalism is suggested by Hume's (1975, 282–84) "sensible knave," a person who recognizes that the unjust and dishonest acts he contemplates are wrong, but is completely unmoved by this realization. More recent writers (e.g., Nichols 2002) have suggested that the sensible knave (or, as philosophers often call him, "the amoralist") is more than a philosophical fiction, since clinical psychologists and other mental health professionals have for some time noted the existence of sociopaths or psychopaths, who appear to *know* the difference between right and wrong but quite generally lack motivation to *do* what is right. If this understanding of the psychopath's moral psychology is accurate, internalism looks to be suffering empirical embarrassment.[13]

Internalists have adopted two quite different responses to this challenge, one conceptual and the other empirical. The first relies on conceptual analysis to argue that a person couldn't really believe that an act is wrong if he has no motivation to avoid performing it. For example, Michael Smith claims it is "a conceptual truth that agents who make moral judgements are motivated accordingly, at least absent weakness of the will and the like" (Smith 1994, 66). Philosophers who adopt this strategy recognize that imaginary knaves and real psychopaths may *say* that something is "morally required" or "morally wrong" and that they may be expressing a judgment that they sincerely accept. But if psychopaths are not motivated in the appropriate way, their words do not mean what non-psychopaths mean by these words and the concepts they express with these words are not the **(p.257)** ordinary moral concepts that non-psychopaths use. Therefore psychopaths "do not *really* make moral judgements at all" (Smith 1994, 67).

This strategy only works if ordinary moral concepts require that people who *really* make moral judgments have the appropriate sort of motivation. But there is considerable disagreement in cognitive science about whether and how concepts are structured, and about how we are to determine when something is built into or entailed by a concept (Margolis and Laurence 1999). Indeed, one widely discussed approach maintains that concepts have no semantically relevant internal structure to be analyzed—thus there are no conceptual entailments (Fodor 1998). Obviously, internalists who appeal to conceptual analysis must reject this account, and in so doing they must take a stand in the broadly empirical debate about the nature of concepts.

Smith is one moral theorist who has taken such a stand. Following Lewis (1970, 1972), Jackson (1994), and others, Smith proposes that a concept can be analyzed by specifying the "maximal consistent set of platitudes" in which the concept is invoked; it is by "coming to treat those platitudes as platitudinous," Smith (1994, 31) maintains, that "we come to have mastery of that concept." If this is correct, the conceptual analysis defense of internalism requires that the maximally consistent set of platitudes invoking the notion of a moral judgment includes a claim to the effect that "agents who make moral judgements are motivated accordingly." Once again, this is an empirical claim. Smith appeals to his own intuitions in its support, but it is of course rather likely that opponents of internalism do not share Smith's intuitions, and it is difficult to say whose intuitions should trump.

In the interests of developing a non-partisan analysis, Nichols (2002) has been running a series of experiments in which philosophically unsophisticated undergraduates are presented with questions like these:

> John is a psychopathic criminal. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. John has hurt, and indeed killed, other people when he has wanted to steal their money. He says that he knows that hurting others is wrong, but that he just doesn't care if he does things that are wrong. Does John really understand that hurting others is morally wrong?
>
> Bill is a mathematician. He is an adult of normal intelligence, but he has no emotional reaction to hurting other people. Nonetheless, Bill never hurts other people simply because he thinks that it is irrational to hurt others. He thinks that any rational person would be like him and not hurt other people. Does Bill really understand that hurting others is morally wrong? (Nichols 2004, 74)

Nichols's preliminary results are exactly the opposite of what Smith would have one expect. An overwhelming majority of subjects maintained that John, the psychopath, did understand that hurting others is morally wrong, while a slight

majority maintained that Bill, the rational mathematician, did not. The implication seems to be that the subjects' **(p.258)** concept of moral judgment does not typically include a "motivational platitude." These results do not, of course, constitute a decisive refutation of Smith's conceptual analysis, since Smith can reply that responses like those Nichols reports would not be part of the maximally consistent set of platitudes that people would endorse after due reflection. But this too is an empirical claim; if Smith is to offer a compelling defense of it he should—with our enthusiastic encouragement—adduce some systematic empirical evidence.

A second internalist strategy for dealing with the problem posed by the amoralist is empirical: even if amoralists are conceptually possible, the internalist may insist, their existence is psychologically impossible. As a matter of psychological fact, this argument goes, people's moral judgments are accompanied by the appropriate sort of motivation.[14] A Kantian elaboration of this idea, on which we will focus, maintains that people's moral judgments are accompanied by the appropriate sort of motivation *unless their rational faculties are impaired*. (We'll shortly see that much turns on the fate of the italicized clause.) Recent papers by Roskies (2003) and Nichols (2002) set out important challenges to this strategy.

Roskies's argument relies on Damasio and colleagues' work with patients suffering injuries to the ventromedial (VM) cortex (Damasio, Tranel, and Damasio 1990; Saver and Damasio 1991; Bechara, Damasio, and Damasio 2000). On a wide range of standard psychological tests, including tests for intelligence and reasoning abilities, these patients appear quite normal. They also do as well as normal subjects on Kohlberg's tests of *moral* reasoning, and when presented with hypothetical situations they offer moral judgments that concur with those of normal subjects. However, these patients appear to have great difficulty acting in accordance with those judgments. As a result, although they often led exemplary lives prior to their injury, their post-trauma social lives are a shambles. They disregard social conventions, make disastrous business and personal decisions, and often engage in anti-social behavior. Accordingly, Damasio and his colleagues describe the VM patients' condition as "acquired sociopathy" (Saver and Damasio 1991).

Roskies maintains that VM patients do not act on their moral judgments because they suffer a *motivational* deficit. Moreover, the evidence indicates that these individuals do not have a *general* difficulty in acting on evaluative judgments; rather, Roskies (2003) maintains, action with respect to moral and social evaluation is differentially impaired. In addition to the behavioral evidence, this interpretation is supported by the anomalous pattern of skin-conductance responses (SCRs) that VM patients display.[15] Normal individuals produce an SCR when presented with emotionally charged or value-laden stimuli, while VM patients typically do not produce SCRs in response to such stimuli. SCRs are not

entirely lacking in VM patients, however. SCRs are produced when VM patients are surprised or startled, for example, demonstrating that the physiological basis for these **(p.259)** responses is intact. In addition, their presence is reliably correlated with cases in which patients' actions are consistent with their judgments about what to do, and their absence is reliably correlated with cases in which patients fail to act in accordance with their judgments. Thus, Roskies contends, the SCR is a reliable indicator of motivation. So the fact that VM patients, unlike normal subjects, do not exhibit SCRs in response to morally charged stimuli suggests that their failure to act in morally charged situations results from a motivational deficit.

On the face of it, acquired sociopathy confounds internalists maintaining that the moral judgments of rational people are, as a matter of psychological fact, always accompanied by appropriate motivation.[16] Testing indicates that the general reasoning abilities of these patients are not impaired, and even their moral reasoning seems to be quite normal. So none of the empirical evidence suggests the presence of a cognitive disability. An internalist might insist that these post-injury judgments are not *genuine* instances of moral judgments because VM patients no longer know the standard meaning of the moral words they use. But unless it is supported by an appeal to a conceptual analysis of the sort we criticized earlier, this is a rather implausible move; as Roskies notes, all tests of VM patients indicate that their language, their declarative knowledge structures, and their cognitive functioning are intact. There are, of course, many questions about acquired sociopathy that remain unanswered and much work is yet to be done. However these questions get answered, the literature on VM patients is one that moral philosophers embroiled in the internalism debate would be ill advised to ignore; once again, the outcome of a debate in ethical theory looks to be contingent on empirical issues.

The same point holds for other work on anti-social behavior. Drawing on Blair's (1995) studies of psychopathic murderers imprisoned in Great Britain, Nichols (2002) has recently argued that the phenomenon of psychopathy poses a deep and complex challenge for internalism. Again, the general difficulty is that psychopaths seem to be living instantiations of Hume's sensible knave: although they appear to be rational and can be quite intelligent, psychopaths are manipulative, remorseless, and devoid of other-regarding concern. While psychopaths sometimes acknowledge that their treatment of other people is wrong, they are quite indifferent about the harm that they have caused; they seem to have no motivation to avoid hurting others (R. D. Hare 1993).

Blair's (1995) evidence complicates this familiar story. He found that psychopaths exhibit surprising deficits on various tasks where subjects are presented with descriptions of "moral" transgressions like a child hitting another child and "conventional" transgressions like a child leaving the classroom without the teacher's permission. From early childhood, normal

children distinguish moral from conventional transgressions on a number of dimensions: they view moral transgressions as more serious, they explain why the acts are wrong by appeal to different factors (harm and fairness for moral transgressions, social **(p.260)** acceptability for conventional transgressions), and they understand conventional transgressions, unlike moral transgressions, to be dependent on authority (Turiel, Killen, and Hedwig 1987; Nucci 1986).

For example, presented with a hypothetical case where a teacher says there is no rule about leaving the classroom without permission, children think it is OK to leave without permission. But presented with a hypothetical where a teacher says there is no rule against hitting other children, children do not judge that hitting is acceptable. Blair has shown that while autistic children, children with Down syndrome, and a control group of incarcerated non-psychopath murderers have relatively little trouble in drawing the moral-conventional distinction and classifying cases along these lines, incarcerated psychopaths are unable to do so.

This inability might be evidence for the hypothesis that psychopaths have a reasoning deficit, and therefore do not pose a problem for internalists who maintain that a properly functioning reasoning faculty reliably generates some motivation to do what one believes one ought to do. But, as Nichols (2002) has pointed out, the issue cannot be so easily resolved, because psychopaths have also been shown to have *affective* responses that are quite different from those of normal subjects. When shown distressing stimuli (like slides of people with dreadful injuries) and threatening stimuli (like slides of an angry man wielding a weapon), normal subjects exhibit much the same suite of physiological responses. Psychopaths, by contrast, exhibit normal physiological responses to threatening stimuli, but abnormally low physiological responses to distressing stimuli (Blair et al. 1997). Thus, Nichols argues, it may well be that the psychopath's deficit is not an abnormal reasoning system, but an abnormal affect system, and it is these affective abnormalities, rather than any rational disabilities, that are implicated in psychopaths' failure to draw the moral-conventional distinction.[17] If his interpretation is correct, it looks as though the existence of psychopaths does undermine the Kantian internalist's empirical generalization: contra the Kantian, there exists a substantial class of individuals *without rational disabilities* who are not motivated by their moral judgments.

We are sympathetic to Nichols's account, but as in the case of VM patients, the internalist is free to insist that a fuller understanding of psychopathy will reveal that the syndrome does indeed involve rational disabilities. Resolving this debate will require conceptual work on how to draw the boundary between reason and affect, and on what counts as an abnormality in each of these domains. But it will also require much more empirical work aimed at understanding exactly how psychopaths and non-psychopaths differ. The internalist—or at least the Kantian internalist—who wishes to diffuse the difficulty posed by psychopathy must

proffer an empirically tenable account of the psychopath's cognitive architecture that locates the posited rational disability. We doubt that **(p.261)** such an account is forthcoming. But—to instantiate once more our take-home message—our present point is that if internalists are to develop such an account, they must engage the empirical literature.

4. Moral Disagreement

Numerous contemporary philosophers, including Brandt (1959), Harman (1977, 125–36), Railton (1986a, b), and Lewis (1989), have proposed dispositional theories of moral rightness or non-moral good, which "make matters of value depend on the affective dispositions of agents" (see Darwall et al. 1997, 28–29).[18] The various versions differ in detail,[19] but a rendering by Brandt is particularly instructive. According to Brandt (1959, 241–70), ethical justification is a process whereby initial judgments about particular cases and general moral principles are revised by testing these judgments against the attitudes, feelings, or emotions that would emerge under appropriately idealized circumstances. Of special importance on Brandt's (1959, 249–51, 261–64) view are what he calls "qualified attitudes"—the attitudes people would have if they were, *inter alia*, (1) impartial, (2) fully informed about and vividly aware of the relevant facts, and (3) free from any "abnormal" states of mind, like insanity, fatigue, or depression.[20]

As Brandt (1959, 281–84) noted, much depends on whether all people would have the same attitudes in ideal circumstances—i.e., on whether their attitudes would *converge* in ideal circumstances. If they would, then certain moral judgments—those where the idealized convergence obtains—are justified for all people, and others—those where such convergence fails to obtain—are not so justified. But if people's attitudes generally fail to converge under idealized circumstances, qualified attitude theory apparently lapses into a version of relativism, since any given moral judgment may comport with the qualified attitudes of one person, and thus be justified for him, while an incompatible judgment may comport with the attitudes of another person, and thus be justified for her.[21]

Brandt, who was a pioneer in the effort to integrate ethical theory and the social sciences, looked primarily to anthropology to help determine whether moral attitudes can **(p.262)** be expected to converge under idealized circumstances. It is of course well known that anthropology includes a substantial body of work, such as the classic studies of Westermarck (1906) and Sumner (1934), detailing the radically divergent moral outlooks found in cultures around the world. But as Brandt (1959, 283–84) recognized, typical ethnographies do not support confident inferences about the convergence of attitudes under *ideal* conditions, in large measure because they often give limited guidance regarding how much of the moral disagreement can be traced to disagreement about factual matters

that are not moral in nature, such as those having to do with religious or cosmological views.

With this sort of difficulty in mind, Brandt (1954) undertook his own anthropological study of Hopi people in the American southwest, and found issues for which there appeared to be serious moral disagreement between typical Hopi and white American attitudes that could not plausibly be attributed to differences in belief about non-moral facts. A notable example is the Hopi attitude towards causing animals to suffer, an attitude that might be expected to disturb many non-Hopis: "[Hopi] children sometimes catch birds and make 'pets' of them. They may be tied to a string, to be taken out and 'played' with. This play is rough, and birds seldom survive long. [According to one informant:] 'Sometimes they get tired and die. Nobody objects to this'" (Brandt 1954, 213).

Brandt (1959, 103) made a concerted effort to determine whether this difference in moral outlook could be traced to disagreement about non-moral facts, but he could find no plausible explanation of this kind; his Hopi informants didn't believe that animals lack the capacity to feel pain, for example, nor did they believe that animals are rewarded for martyrdom in the afterlife. According to Brandt (1954, 245), the Hopi do not regard animals as unconscious or insensitive; indeed, they apparently regard animals as "closer to the human species than does the average white man." The best explanation of the divergent moral judgments, Brandt (1954, 245) concluded, is a "basic difference of attitude." Accordingly, although he cautions that the uncertainties of ethnography make confident conclusions on this point difficult, Brandt (1959, 284) argues that accounts of moral justification like his qualified attitude theory *do* end in relativism, since "groups do sometimes make divergent appraisals when they have identical beliefs about the objects."

Of course, the observation that persistent moral disagreement appears to problematize moral argument and justification is not unique to Brandt. While the difficulty is long familiar, contemporary philosophical discussion was spurred by Mackie's (1977, 36–38) "argument from relativity" or, as it is called by later writers, the "argument from disagreement" (Brink 1989, 197; Loeb 1998). Such "radical" differences in moral judgment as are frequently observed, Mackie (1977, 36) argued, "make it difficult to treat those judgments as apprehensions of objective truths." As we see it, the problem is not only that moral disagreement often persists, but also that for important instances of moral disagreement—such as the treatment of animals—it is obscure what sort of considerations, be they methodological or substantive, *could* settle the issues (see Sturgeon 1988, 229). Indeed, moral disagreement might be plausibly expected to continue even when the disputants are in methodological agreement concerning the appropriate standards for moral **(p.263)** argument. One way of putting the point is to say that application of the same method may, for different individuals or cultures, yield divergent moral judgments that are equally

acceptable by the lights of the method, even in reflective conditions that the method countenances as ideal.[22]

In contemporary ethical theory, an impressive group of philosophers are “moral realists” (see Railton 1986a, b; Boyd 1988; Sturgeon 1988; Brink 1989; M. Smith 1994). Adherents to a single philosophical creed often manifest doctrinal differences, and that is doubtless the case here, but it is probably fair to say that most moral realists mean to resist the argument from disagreement and reject its relativist conclusion. For instance, Smith’s (1994, 9; cf. 13) moral realism requires the objectivity of moral judgment, where objectivity is construed as “the idea that moral questions have correct answers, that the correct answers are made correct by objective moral facts, that moral facts are determined by circumstances, and that, by engaging in moral argument, we can discover what these objective moral facts are.” There’s a lot of philosophy packed into this statement, but it looks as though Smith is committed to the thought, contra our relativist, that moral argument, or at least moral argument of the right sort, can settle moral disagreements. Indeed, for Smith (1994, 6), the notion of objectivity “signifies the possibility of a convergence in moral views,” so the prospects for his version of moral realism depend on the argument from disagreement not going through.[23] But can realists like Smith bank on the argument’s failure?

Realists may argue that, in contrast to the impression one gets from the anthropological literature, there already exists substantial moral convergence. But while moral realists have often taken pretty optimistic positions on the extent of actual moral agreement (e.g., Sturgeon 1988, 229; M. Smith 1994, 188), there is no denying that there is an abundance of persistent moral disagreement. That is, on many moral issues—think of abortion and capital punishment—there is a striking failure of convergence even after protracted argument. The relativist has a ready explanation for this phenomenon: moral judgment is not objective in Smith’s sense, and moral argument cannot be expected to accomplish what Smith and other realists think it can.[24] Conversely, the realist’s task is to *explain away* **(p.264)** failures of convergence; she must provide an explanation of the phenomena consistent with it being the case that moral judgment is objective and moral argument is rationally resolvable. For our purposes, what needs to be emphasized is that the relative merits of these competing explanations cannot be fairly determined without close discussion of actual cases. Indeed, as acute commentators with both realist (Sturgeon 1988, 230) and anti-realist (Loeb 1998, 284) sympathies have noted, the argument from disagreement cannot be evaluated by a priori philosophical means alone; what’s needed, as Loeb observes, is “a great deal of further empirical research into the circumstances and beliefs of various cultures.”

Brandt (1959, 101–2) lamented that the anthropological literature of his day did not always provide as much information on the exact contours and origins of moral attitudes and beliefs as philosophers wondering about the prospects for

convergence might like. However, social psychology and cognitive science have recently produced research which promises to further discussion; the closing decades of the twentieth century witnessed an explosion of “cultural psychology” investigating the cognitive and emotional processes of different cultures (Shweder and Bourne 1982; Markus and Kitayama 1991; Ellsworth 1994; Nisbett and Cohen 1996; Nisbett 1998; Kitayama and Markus 1999). A representative finding is that East Asians are more sensitive than Westerners to the field or context as opposed to the object or actor in their explanations of physical and social phenomena, a difference that may be reflected in their habits of ethical judgment. Here we will focus on some cultural differences found rather closer to home, differences discovered by Nisbett and his colleagues while investigating regional patterns of violence in the American North and South. We argue that these findings support Brandt’s pessimistic conclusions regarding the possibility of convergence in moral judgment.

The Nisbett group’s research can be seen as applying the tools of cognitive social psychology to the “culture of honor,” a phenomenon that anthropologists have documented in a variety of groups around the world. Although such peoples differ in many respects, they manifest important commonalties:

> A key aspect of the culture of honor is the importance placed on the insult and the necessity to respond to it. An insult implies that the target is weak enough to be bullied. Since a reputation for strength is of the essence in the culture of honor, the individual who insults someone must be forced to retract; if the instigator refuses, he must be punished—with violence or even death. (Nisbett and Cohen 1996, 5)

According to Nisbett and Cohen (1996, 5–9), an important factor in the genesis of southern honor culture was the presence of a herding economy. Apparently, honor cultures are particularly likely to develop where resources are liable to theft, and where the state’s coercive apparatus cannot be relied upon to prevent or punish **(p.265)** thievery. These conditions often occur in relatively remote areas where herding is the main viable form of agriculture; the “portability” of herd animals makes them prone to theft. In areas where farming rather than herding is the principal form of subsistence, cooperation among neighbors is more important, stronger government infrastructures are more common, and resources—like decidedly unportable farmland—are harder to steal. In such agrarian social economies, cultures of honor tend not to develop. The American South was originally settled primarily by peoples from remote areas of Britain. Since their homelands were generally unsuitable for farming, these peoples have historically been herders; when they emigrated from Britain to the South, they initially sought out remote regions suitable for herding, and in such regions, the culture of honor flourished.

In the contemporary South police and other government services are widely available and herding has all but disappeared as a way of life, but certain sorts of violence continue to be more common than they are in the North. Nisbett and Cohen (1996) maintain that patterns of violence in the South, as well as attitudes towards violence, insults, and affronts to honor, are best explained by the hypothesis that a culture of honor persists among contemporary white non-Hispanic southerners. In support of this hypothesis, they offer a compelling array of evidence, including:

- demographic data indicating that (1) among southern whites, homicide rates are higher in regions more suited to herding than agriculture, and (2) white males in the South are much more likely than white males in other regions to be involved in homicides resulting from arguments, although they are *not* more likely to be involved in homicides that occur in the course of a robbery or other felony (Nisbett and Cohen 1996, ch. 2);
- survey data indicating that white southerners are more likely than northerners to believe that violence would be "extremely justified" in response to a variety of affronts, and that if a man failed to respond violently, he was "not much of a man" (Nisbett and Cohen 1996, ch. 3);
- legal scholarship indicating that southern states "give citizens more freedom to use violence in defending themselves, their homes, and their property" than do northern states (Nisbett and Cohen 1996, 63).

Two experimental studies—one in the field, the other in the laboratory—are especially striking.

In the field study (Nisbett and Cohen 1996, 73–75), letters of inquiry were sent to hundreds of employers around the United States. The letters purported to be from a hardworking 27-year-old Michigan man who had a single blemish on his otherwise solid record. In one version, the "applicant" revealed that he had been convicted for manslaughter. The applicant explained that he had been in a fight with a man who confronted him in a bar and told onlookers that "he and my fiancée were sleeping together. He laughed at **(p.266)** me to my face and asked me to step outside if I was man enough." According to the letter, the applicant's nemesis was killed in the ensuing fray. In the other version of the letter, the applicant revealed that he had been convicted of motor vehicle theft, perpetrated at a time when he needed money for his family. Nisbett and his colleagues assessed 112 letters of response, and found that southern employers were significantly more likely to be cooperative and sympathetic in response to the manslaughter letter than were northern employers, while no regional differences were found in responses to the theft letter. One southern employer responded to the manslaughter letter as follows (Nisbett and Cohen 1996, 75):

> As for your problems of the past, anyone could probably be in the situation you were in. It was just an unfortunate incident that shouldn't be held against you. Your honesty shows that you are sincere.... I wish you the best of luck for your future. You have a positive attitude and a willingness to work. These are qualities that businesses look for in employees. Once you are settled, if you are near here, please stop in and see us.

No letters from northern employers were comparably sympathetic.

In the laboratory study (Nisbett and Cohen 1996, 45–48) subjects—white males from both northern and southern states attending the University of Michigan—were told that saliva samples would be collected to measure blood sugar as they performed various tasks. After an initial sample was collected, the unsuspecting subject walked down a narrow corridor where an experimental confederate was pretending to work on some filing. Feigning annoyance at the interruption, the confederate bumped the subject and called him an "asshole." A few minutes after the incident, saliva samples were collected and analyzed to determine the level of cortisol—a hormone associated with high levels of stress, anxiety, and arousal, and testosterone—a hormone associated with aggression and dominance behavior. As Figure 11.1 indicates, southern subjects showed dramatic increases in cortisol and testosterone levels, while northerners exhibited much smaller changes.

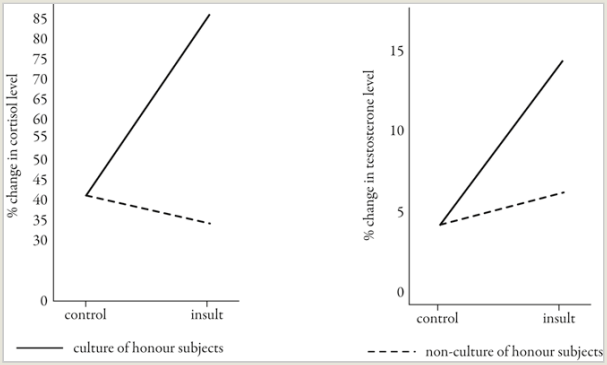

Figure 11.1 The results of an experiment by Nisbett and Cohen in which levels of cortisol and testosterone increased much more substantially in culture of honor subjects who were insulted by a confederate

The two studies just described suggest that southerners respond more strongly to insult than northerners, and take a more sympathetic view of others who do so, manifesting just the sort of attitudes that are supposed to typify honor cultures. We think that the data assembled by Nisbett and his colleagues make a persuasive case that a culture of honor persists in the American South. Apparently, this culture affects people's judgments, attitudes, emotions, behavior, and even their physiological responses. Additionally, there is evidence that child-rearing practices play a significant role in passing the culture of honor on from one generation to the next, and also that relatively permissive laws regarding gun ownership, self-defense, and corporal punishment in the schools both reflect and reinforce southern honor culture (Nisbett and Cohen 1996, 60–

63, 67–69). In short, it seems to us that the culture of honor is deeply entrenched in contemporary southern **(p.267)** culture, despite the fact that many of the material and economic conditions giving rise to it no longer widely obtain.[25]

We believe that the North-South cultural differences adduced by Nisbett and colleagues support Brandt's conclusion that moral attitudes will often fail to converge, even under ideal conditions. The data should be especially troubling for the realist, for despite the differences that we have been recounting, contemporary northern and southern Americans might be expected to have rather more in common—from circumstance to language to belief to ideology—than do, say, Yanomamö and Parisians. So if there is little ground for expecting convergence under ideal conditions in the case at hand, there is probably little ground in a good many others. To develop our argument a bit further, let us revisit the idealization conditions mentioned at the beginning of this section: impartiality, full factual information, and normality.

Impartiality. One strategy favored by moral realists concerned to explain away moral disagreement is to say that such disagreement stems from the distorting effects of individual interest (see Sturgeon 1988, 229–30); perhaps persistent disagreement doesn't so much betray deep features of moral argument and judgment as it does the doggedness with which individuals pursue their perceived advantage. For instance, seemingly moral disputes over the distribution of wealth may be due to perceptions—perhaps mostly inchoate—of individual and class interests rather than to principled disagreement about **(p. 268)** justice; persisting moral disagreement in such circumstances fails the impartiality condition, and is therefore untroubling to the moral realist.

But it is rather implausible to suggest that North-South disagreements over when violence is justified will fail the impartiality condition. There is no reason to think that southerners would be unwilling to universalize their judgments across relevantly similar individuals in relevantly similar circumstances, as indeed Nisbett and Cohen's "letter study" suggests. One can advocate a violent honor code without going in for special pleading.[26] We do not intend to denigrate southern values; our point is that while there may be good reasons for criticizing the honor-bound southerner, it is not obvious that the reason can be failure of impartiality, if impartiality is (roughly) to be understood along the lines of a willingness to universalize one's moral judgments.

Full and vivid awareness of relevant non-moral facts. Moral realists have argued that moral disagreements very often derive from disagreement about non-moral issues. According to Boyd (1988, 213; cf. Brink 1989, 202–3; Sturgeon 1988, 229), "careful philosophical examination will reveal … that agreement on nonmoral issues would eliminate *almost all* disagreement about the sorts of moral issues which arise in ordinary moral practice." Is this a plausible

conjecture for the data we have just considered? We find it hard to imagine what agreement on non-moral facts could do the trick, for we can readily imagine that northerners and southerners might be in full agreement on the relevant non-moral facts in the cases described. Members of both groups would presumably agree that the job applicant was cuckolded, for example, or that calling someone an “asshole” is an insult. We think it much more plausible to suppose that the disagreement resides in differing and deeply entrenched evaluative attitudes regarding appropriate responses to cuckolding, challenge, and insult.

Savvy philosophical readers will be quick to observe that terms like “challenge” and “insult” look like “thick” ethical terms, where the evaluative and descriptive are commingled (see Williams 1985, 128–30); therefore, it is very difficult to say what the extent of the factual disagreement is. But this is of little help for the expedient under consideration, since the disagreement-in-non-moral-fact response apparently *requires* that one *can* disentangle factual and moral disagreement.

It is of course possible that full and vivid awareness of the non-moral facts might motivate the sort of change in southern attitudes envisaged by the (at least the northern) moral realist; were southerners to become vividly aware that their culture of honor was implicated in violence, they might be moved to change their moral outlook. (We take this way of putting the example to be the most natural one, but nothing philosophical turns on it. If you like, substitute the possibility of bloody-minded northerners endorsing **(p.269)** honor values after exposure to the facts.) On the other hand, southerners might insist that the values of honor should be nurtured even at the cost of promoting violence; the motto “Death before dishonor,” after all, has a long and honorable history. The burden of argument, we think, lies with the realist who asserts—culture and history notwithstanding—that southerners would change their mind if vividly aware of the pertinent facts.

Freedom from abnormality. Realists may contend that much moral disagreement may result from failures of rationality on the part of discussants (Brink 1989, 199–200). Obviously, disagreement stemming from cognitive impairments is no embarrassment for moral realism; at the limit, that a disagreement persists when some or all disputing parties are quite insane shows nothing deep about morality. But it doesn’t seem plausible that southerners’ more lenient attitudes towards certain forms of violence are readily attributed to widespread cognitive disability. Of course, this is an empirical issue, and we don’t know of any evidence suggesting that southerners suffer some cognitive impairment that prevents them from understanding demographic and attitudinal factors in the genesis of violence, or any other matter of fact. What is needed to press home a charge of irrationality is evidence of cognitive impairment independent of the attitudinal differences, and further evidence that this impairment is implicated in adherence to the disputed values in the face of the (putatively) undisputed

non-moral facts. In this instance, as in many others, we have difficulty seeing how charges of abnormality or irrationality can be made without one side begging the question against the other.

We are inclined to think that Nisbett and colleagues' work represents a potent counter-example to any theory maintaining that rational argument tends to convergence on important moral issues; the evidence suggests that the North-South differences in attitudes towards violence and honor might well persist even under the sort of ideal conditions we have considered. Admittedly, our conclusions must be tentative. On the philosophical side, we have not considered every plausible strategy for "explaining away" moral disagreement and grounding expectations of convergence.[27] On the empirical side, we have reported on but a few studies, and those we do consider here, like any empirical work, might be criticized on either conceptual or methodological grounds.[28] Finally, we should make clear what we are *not* claiming: we do not take our conclusions here—even if fairly earned—to be a "refutation" of moral realism, in as much as there may be versions of moral realism that do not require convergence. Rather, we hope to have given an idea of the empirical work philosophers must encounter if they are to make defensible conjectures regarding moral disagreement. Our theme recurs: Responsible **(p.270)** treatment of the empirical issues requires reference to empirical science, whatever the science is ultimately taken to show.

5. Thought Experiments

Ethical reflection is often held to involve comparing general principles and responses to particular cases; commitment to a principle may compel the renunciation of a particular response, or commitment to a particular response may compel modification or renunciation of a general principle (Brandt 1959, 244–52; Rawls 1971, 20–21, 49). This emphasis on particular cases is not peculiar to ethics: "intuition pumps" or "thought experiments" have long been central elements of philosophical method (Dennett 1984, 17–18). In the instances we consider here, a thought experiment presents an example, typically a hypothetical example, in order to elicit some philosophically telling response; if a thought experiment is successful, it may be concluded that competing theories must account for the resulting response.[29] To extend the imagery of experimentation, responses to thought experiments are supposed to serve an evidential role in philosophical theory choice; the responses are data competing theories must accommodate.[30]

In ethics, one—we do not say the only—familiar rendering of the methodology is this: if an audience's ethical responses to a thought experiment can be expected to conflict with the response a theory prescribes for the case, the theory has suffered a counter-example. For instance, it is often claimed that utilitarian prescriptions for particular cases will conflict with the ethical responses many people have to those cases (e.g., Williams 1973, 99). The ethics literature is rife

with claims to the effect that “many of us” or “we” would respond in a specified way to a given example, and such claims are often supposed to have philosophical teeth.[31] But who is this “we”? And how do philosophers know what this “we” thinks?

Initially, it doesn’t look like “we” should be interpreted as “we philosophers.” The difficulty is not that this approach threatens a sampling error, although it is certainly true that philosophers form a small and peculiar group. Rather, the problem is that philosophers can be expected to respond to thought experiments in ways that reflect their theoretical **(p.271)** predilections: utilitarians’ responses to a thought experiment might be expected to plump for maximizing welfare, integrity and loyalty be damned, while the responses of Aristotelians and Kantians might plump in the opposite direction. If so, the thought experiment can hardly be expected to *resolve* the debate, since philosophers’ responses to the example are likely to *reflect* their position in the debate.

The audience of appeal often seems to be some variant of “ordinary folk” (see Jackson 1998, 118, 129; Jackson and Pettit 1995, 22–29; Lewis 1989, 126–29). Of course, the relevant folk must possess such cognitive attainments as are required to understand the case at issue; very young children are probably not an ideal audience for thought experiments. Some philosophers may want to insist that the relevant responses are the “considered judgments” or “reflective intuitions” of people with the training required to see “what is philosophically at stake.” But there is peril in insisting that the relevant cognitive attainments be some sort of “philosophical sophistication.” Once again, if the responses are to help adjudicate between competing theories, the responders must be more or less theoretically neutral, but this sort of neutrality, we suspect, is rather likely to be vitiated by philosophical education. (Incredibly enough, informal surveys suggest that *our* students are overwhelmingly ethical naturalists!)

However exactly the philosophically relevant audience is specified, there are empirical questions that must be addressed in determining the philosophical potency of a thought experiment. In science, not all experiments produce data of evidentiary value; sampling errors and the failure of experimental designs to effectively isolate variables are two familiar ways in which experiments go wrong. Data resulting from such experiments are tainted, or without evidential value; analogously, in evaluating responses to a thought experiment, one needs to consider the possibility of taint. In particular, when deciding what philosophical weight to give a response to a thought experiment, philosophers need to determine the origins of the response. What features of the example are implicated in a response—are people responding to the substance of the case, or the style of exposition? What features of the audience are implicated in a response—do different demographic groups respond to an example differently? Such questions raise the following concern: ethical responses to thought experiments may be strongly influenced by ethically irrelevant characteristics of

example and audience. Whether a characteristic is ethically relevant is a matter for philosophical discussion, but determining the status of a particular thought experiment also requires empirical investigation of its causally relevant characteristics; responsible philosophical discussion cannot rely on guesswork in this regard. We shall now give two examples illustrating our concerns about tainted origins, one corresponding to each of the two questions just asked.

Tversky and Kahneman presented subjects with the following problem:

> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the **(p.272)** disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die. (Tversky and Kahneman 1981, 453)

On the first version of the problem most subjects thought that Program A should be adopted. But on the second version most chose Program D, despite the fact that the outcome described in A is identical to the one described in C. The disconcerting implication of this study is that ethical responses may be strongly influenced by the manner in which cases are described or framed. Many effects of framing differences, such as that between 200 of 600 people being saved and 400 of 600 dying, we are strongly inclined to think, are ethically irrelevant influences on ethical responses (compare Horowitz 1998; Sinnott-Armstrong 2005). Unless this sort of possibility can be confidently eliminated, one should hesitate to rely on responses to a thought experiment for adjudicating theoretical controversies. Again, such possibilities can only be eliminated through systematic empirical work.[32]

Audience characteristics may also affect the outcome of thought experiments. Haidt and associates (Haidt, Koller, and Dias 1993, 613) presented stories about “harmless yet offensive violations of strong social norms” to men and women of high and low socio-economic status (SES) in Philadelphia (USA), Porto Alegre, and Recife (both in Brazil). For example: (“A man goes to the supermarket once a week and buys a dead chicken. But before cooking the chicken, he has sexual

intercourse with it. Then he cooks it and eats it") (Haidt et al. 1993, 617). Lower SES subjects tended to "moralize" harmless and offensive behaviors like that in the chicken story: these subjects were more inclined than their privileged counterparts to say that the actor should be "stopped or punished," and more inclined to deny that such behaviors would be "OK" if customary in a given country (Haidt et al. 1993, 618–19). The point is not that lower SES subjects are mistaken in their moralization of such behaviors while the urbanity of higher SES subjects represents the **(p.273)** most rationally defensible response. To recall our previous discussion of moral disagreement, the difficulty is deciding which of the conflicting responses to privilege, when both sorts of responses may be the function of more or less arbitrary cultural factors.

In presenting the Haidt group's work to philosophical audiences, our impression is that they typically decline to moralize the offensive behaviors, and we ourselves share their tolerant attitude. But of course philosophical audiences—by virtue of educational attainments if not stock portfolios—are overwhelmingly high SES. Haidt's work suggests that it is a mistake for a philosopher to say, as Jackson (1998, 32 ff.; cf. 37) does, that "my intuitions reveal the folk conception in as much as I am reasonably entitled, as I usually am, to regard myself as typical." The question is: Typical of what demographic? Are philosophers' ethical responses to thought experiments determined by the philosophical substance of the examples, or by cultural idiosyncrasies that are very plausibly thought to be ethically irrelevant? Once again, until such possibilities are ruled out by systematic empirical investigation, the philosophical heft of a thought experiment is open to question.[33]

The studies just described raise provocative questions about *how* responses to thought experiments are generated, but there may be equally provocative questions about *what* responses people actually have. And, to sound our now familiar theme, this question is one not credibly answered by guesswork. Indeed, we suspect that philosophical speculations about what responses to thought experiments are conventional may be wrong surprisingly often. We'll now report on one study conducive to such suspicions.

One of the most famous of philosophical conundrums, that of determinism and responsibility, can be derived—on one way of formulating the difficulty—from the juxtaposition of three claims that are individually quite plausible, but seem impossible to hold jointly:

(MRT) *Moral responsibility thesis*:	Human beings are sometimes morally responsible for their behavior.

(CT) *Causal thesis*:	All human behavior is linked to antecedent events by deterministic causal laws. (See Scanlon 1988, 152.)
(PAP) *Principle of alternate possibilities*:	A "person is morally responsible for what he has done only if he could have done otherwise." (See Frankfurt 1988, 1.)

(p.274) Here's one way of putting it: If CT is true, it looks as though it is never the case that people could have done otherwise, but then, given PAP, MRT must be false.[34] There are three standard responses to this trilemma. Two sorts of incompatibilists hold that MRT and CT cannot be held simultaneously: hard determinists (see Smart 1961, 303–6) reject MRT,[35] while libertarians (e.g., Kane 1996) insist that CT admits of exceptions in the case of human behavior, and are thus able to maintain MRT. Compatibilists, on the other hand, assert that MRT and CT can be simultaneously maintained; one well-known expedient is to reject PAP, and insist that people may be legitimately held responsible even when they could not have done otherwise (see Frankfurt 1988, 1–12).

The literature is voluminous, and the proffered solutions range from controversial to deeply unsatisfying; indeed, there is heated disagreement as to what exactly the problem is (Dennett 1984, 1–19). Discretion being the best part of valor, we won't review the arguments here. Given our present concerns, we instead consider objections to the effect that compatibilism is in some sense badly counter-intuitive. One way of forming this complaint is to say that people's "reactive attitudes"—ethical responses like anger, resentment, guilt, approbation, admiration, and the like—manifest a commitment to incompatibilism.[36] Here is Galen Strawson (1986, 88) on what he calls the "incompatibilist intuition":

> The fact that the incompatibilist intuition has such power for us is as much a natural fact about cogitative beings like ourselves as is the fact of our quite unreflective commitment to the reactive attitudes. What is more, the roots of the incompatibilist intuition lie deep in the … reactive attitudes…. The reactive attitudes enshrine the incompatibilist intuition.[37]

Let's do a little unpacking. On Strawson's (1986, 31; cf. 2, 84–88) rendering, incompatibilism is the view that the falsity of determinism is a necessary condition for moral responsibility. To suggest that the "incompatibilist intuition" is widespread, then, may be thought to imply that people's (possibly tacit) body of moral beliefs includes commitment to the **(p.275)** claim that CT is incompatible with MRT.[38] This is an empirical claim. Moreover, it is an empirical claim that looks to entail predictions about people's moral responses. What are the responses in question?

Like many other philosophers making empirical claims about human cognition and behavior, Strawson says relatively little about what predictions he thinks his claims entail. We won't put predictions in Strawson's mouth; instead, we'll consider one prediction that looks to follow from positing an incompatibilist intuition, at least on the familiar rendering of incompatibilism we've followed. Attributing a widespread commitment to an incompatibilist intuition is plausibly thought to involve the following prediction: for cases where the actor is judged unable to have done otherwise, people will not hold the actor responsible for what she has done.[39] In as much as this prediction is a good one, people should respond to thought experiments depicting an actor unable to do otherwise by abjuring attributions of responsibility and the associated reactive attitudes.

In a compatibilist spirit inspired by the work of Harry Frankfurt (1988), Woolfolk, Doris, and Darley (2006) hypothesized that observers may hold actors responsible even when the observers judge that the actors could not have done otherwise, at least in cases where the actors appear to manifest "identification." Very roughly, the idea is that the actor is identified with a behavior—and is therefore responsible for it—to the extent she "embraces" the behavior (or its motive), or performs it "wholeheartedly."[40] Woolfolk et al.'s suspicion was, in effect, that people's (possibly tacit) theory of responsibility is, contra Galen Strawson and others, compatibilist.

In one of the Woolfolk et al. studies, subjects read a story about two married couples vacationing together. According to the story, one of the vacationers has discovered that his wife is having an affair with his opposite number in the foursome; on the flight home, the vacationers' plane is hijacked, and armed hijackers order the cuckold to shoot the man who has been having an affair with his wife. In a "low identification" variation, the story contained the following material:

> Bill was horrified. At that moment Bill was certain about his feelings. He did *not* want to kill Frank, even though Frank was his wife's lover. But although he was appalled by the situation and beside himself with distress, he reluctantly placed the pistol at Frank's temple and proceeded to blow his friend's brains out.

(p.276) Conversely, in a "high identification" variation, the embittered cuckold embraces his opportunity:

> Despite the desperate circumstances, Bill understood the situation. He had been presented with the opportunity to kill his wife's lover and get away with it. And at that moment Bill was certain about his feelings. He wanted to kill Frank. Feeling no reluctance, he placed the pistol at Frank's temple and proceeded to blow his friend's brains out.

Consistent with Woolfolk and colleagues' hypothesis, the high-identification actor was judged more responsible, more appropriately blamed, and more properly subject to guilt than the low-identification actor.[41]

It is tempting to conclude that at least for the Woolfolk group's subjects (philosophy and psychology undergraduates at the University of California and Rutgers University), the incompatibilist intuition does not appear to be deeply entrenched. But at this point the incompatibilist will be quick to object: the above study may suggest that responsibility attributions are influenced by identification, but it says nothing about commitment to the incompatibilist intuition, because subjects may not have believed that the actor could not have done otherwise, and subjects therefore cannot be interpreted as attributing responsibility in violation of PAP. People may think that even when coerced, actors "always have a choice"; in the classic "your money or your life" scenario, the person faced with this unpleasant dilemma can always opt for her life. (We hasten to remind anyone tempted in such a bull-headed direction that the disjunct need not be exclusive!)

To address this objection, Woolfolk et al. attempted to elevate perceived constraint to the "could not have done otherwise" threshold:

> The leader of the kidnappers injected Bill's arm with a "compliance drug"—a designer drug similar to sodium pentothal, "truth serum." This drug makes individuals unable to resist the demands of powerful authorities. Its effects are similar to the impact of expertly administered hypnosis; it results in total compliance. To test the effects of the drug, the leader of the kidnappers shouted at Bill to slap himself. To his amazement, Bill observed his own right hand administering an open-handed blow to his own left cheek, although he had no sense of having willed his hand to move. The leader then handed Bill a pistol with one bullet in it. Bill was ordered to shoot Frank in the head…. when Bill's hand and arm moved again, placing the pistol at his friend's temple, Bill had no feeling that he had moved his arm to point the gun; it felt as though the gun had moved itself into position. Bill thought he noticed his finger moving on the trigger, but could **(p.277)** not feel any sensations of movement. While he was observing these events, feeling like a puppet, passively observing his body moving in space, his hand closed on the pistol, discharging it and blowing Frank's brains out.

Strikingly, subjects appeared willing to attribute responsibility to the shooter even here: once again, a high-identification actor was judged more responsible, more appropriately blamed, and more properly subject to guilt than a low-identification actor. No doubt this is not the most "naturalistic" scenario, but neither is it outlandish by philosophical standards. And it certainly looks to be a case where the actor would be perceived to fail the standard for responsibility

set by PAP.[42] Indeed, Woolfolk et al. found that subjects were markedly less likely to agree to statements asserting that the actor "was free to behave other than he did," and "could have behaved differently than he did," than they were in the case of simple coercion described above. These results look to caution against positing a widespread commitment to the incompatibilist intuition. Deciding empirical issues concerning habits of responsibility attribution will not, of course, decide the philosophical dispute between compatibilists and incompatibilists. Yet in so far as the incompatibilist is making claims to the effect that compatibilists cannot accommodate entrenched habits of moral response, the empirical evidence is entirely relevant.

Once more, some philosophers may insist that the responses of interest are not the relatively unschooled or intuitive responses of experimental subjects like the Woolfolk group's undergraduates, but the tutored judgments of philosophers. We've already given some reasons for regarding this strategy with suspicion, but it seems to us especially problematic for the particular case of responsibility. Philosophical arguments about responsibility, it seems to us, often lean heavily on speculation about everyday practice. For example, Peter Strawson's (1982, 64, 68) extremely influential exposition repeatedly stresses the importance of reactive attitudes in "ordinary inter-personal relationships." While it may not be too much of a stretch to imagine that philosophers sometimes indulge in such relationships, it *is* a stretch to suppose that they are the only folk who do so. It is very plausible to argue—as indeed those who have deployed something like the incompatibilist intuition have done—that the contours of the everyday practice of responsibility attribution serve as a (defeasible) constraint on philosophical theories of responsibility: if the theory cannot accommodate the practice, it owes, at a bare minimum, a debunking account of the practice. One might insist that philosophical theorizing about responsibility is not accountable to ordinary practice, but this is to make a substantial break with important elements of the tradition.

There are a couple of ways in which philosophers can avoid the sorts of empirical difficulties we have been considering. First, they can deny that responses to particular cases have evidential weight in ethical theory choice, as some utilitarians—unsurprisingly, **(p.278)** given the rather startling implications of their position—have been inclined to do (e.g., Kagan 1989, 10–15; Singer 2000, xviii). Alternatively, they can appeal to the results of thought experiments in an expository rather than an evidential role; for example, a thought experiment might be used by an author to elucidate her line of reasoning without appealing to the responses of an imagined audience like "many of us." To some philosophers, such solutions will seem rather methodologically draconian, threatening to isolate ethical theory from the experience of ethical life (see Williams 1985, 93–119, esp. 116–19). But our point here is less grand: many users of thought experiments in ethics apparently have been—and we strongly suspect will continue to be—in the business of

forwarding an imagined consensus on their thought experiments as evidence in theory choice. For these philosophers we offer the following methodological prescription: a credible philosophical methodology of thought experiments must be supplemented by a cognitive science of *thought* experiments that involves systematic investigation with *actual* experiments. There are just too many unanswered questions regarding the responses people have, and the processes by which they come to have them. We've no stake in any particular answers to such questions. What we do have a stake in, as we have throughout, is the observation that responsible answers to such questions will be informed by systematic empirical investigation.

6. Conclusion

We needn't linger on goodbyes; the main contours of our exposition should by now be tolerably clear. We have surveyed four central topics in ethical theory where empirical claims are prominent: character, moral motivation, moral disagreement, and thought experiments. We have argued that consideration of work in the biological, behavioral, and social sciences promises substantive philosophical contributions to controversy surrounding such topics as virtue ethics, internalism, moral realism, and moral responsibility. If our arguments are successful, we have also erected a general methodological standard: philosophical ethics can, and indeed must, interface with the human sciences.[43]

References

Bibliography references:

Annas, J. 1993. *The Morality of Happiness*. New York: Oxford University Press.

Anscombe, G. E. M. 1958. Modern Moral Philosophy. *Philosophy* 33: 1–19.

(p.279) Aristotle. 1984. *The Complete Works of Aristotle*. Ed. J. Barnes. Princeton: Princeton University Press.

Athanassoulis, N. 2000. A Response to Harman: Virtue Ethics and Character Traits. *Proceedings of the Aristotelian Society* 100: 215–22.

Audi, R. 1995. Acting from Virtue. *Mind* 104: 449–71.

Baron, J. 1994. Nonconsequentialist Decisions. *Behavioral and Brain Sciences* 17: 1–42.

Baron, J. 2001. *Thinking and Deciding*. 3rd ed. Cambridge: Cambridge University Press.

Bechara, A., H. Damasio, and A. R. Damasio. 2000. Emotion, Decision Making and the Orbitofrontal Cortex. *Cerebral Cortex* 10: 295–307.

Becker, L. C. 1998. *A New Stoicism*. Princeton: Princeton University Press.

Bennett, W. J. 1993. *The Book of Virtues: A Treasury of Great Moral Stories*. New York: Simon and Schuster.

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press.

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

Blair, R. J., L. Jones, F. Clark, and M. Smith. 1997. The Psychopathic Individual: A Lack of Responsiveness to Distress Cues? *Psychophysiology* 34: 192–98.

Blum, L. A. 1994. *Moral Perception and Particularity*. Cambridge: Cambridge University Press.

Bok, H. 1996. Acting Without Choosing. *Noûs* 30: 174–96.

Boyd, R. N. 1988. How to Be a Moral Realist. In C. Sayre-McCord, ed., *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.

Brandt, R. B. 1954. *Hopi Ethics: A Theoretical Analysis*. Chicago: University of Chicago Press.

Brandt, R. B. 1959. *Ethical Theory: The Problems of Normative and Critical Ethics*. Englewood Cliffs, NJ: Prentice-Hall.

Brandt, R. B. 1970. Traits of Character: A Conceptual Analysis. *American Philosophical Quarterly* 7: 23–37.

Bratman, M. E. 1996. Identification, Decision, and Treating as a Reason. *Philosophical Topics* 24: 1–18.

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

Campbell, J. 1999. Can Philosophical Accounts of Altruism Accommodate Experimental Data on Helping Behaviour? *Australasian Journal of Philosophy* 77: 26–45.

Cooper, J. M. 1999. *Reason and Emotion: Essays on Ancient Moral Psychology and Ethical Theory*. Princeton: Princeton University Press.

Damasio, A. R., D. Tranel, and H. Damasio. 1990. Individuals with Sociopathic Behavior Caused by Frontal Damage Fail to Respond Autonomically to Social Stimuli. *Behavioral Brain Research* 41: 81–94.

Daniels, N. 1979. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *Journal of Philosophy* 76: 256–84.

Darley, J. M., and C. D. Batson. 1973. From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior. *Journal of Personality and Social Psychology* 27: 100–108.

D'Arms, J., and D. Jacobson. 2000. Sentiment and Value. *Ethics* 110: 722–48.

Darwall, S. L. 1983. *Impartial Reason*. Ithaca, NY: Cornell University Press.

(p.280) Darwall, S. L. 1989. Moore to Stevenson. In Robert Cavalier, James Gouinlock, and James Sterba, eds., *Ethics in the History of Philosophy*. London: Macmillan.

Darwall, S. L., A. Gibbard, and P. Railton, eds. 1997. *Moral Discourse and Practice: Some Philosophical Approaches*. New York: Oxford University Press.

Deigh, J. 1999. Ethics. In R. Audi, ed., *The Cambridge Dictionary of Philosophy*. Cambridge: Cambridge University Press.

Dennett, D. C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.

Dent, N. J. H. 1975. Virtues and Actions. *Philosophical Quarterly* 25: 318–35.

DePaul, M. 1999. Character Traits, Virtues, and Vices: Are There None? In *Proceedings of the 20th World Congress of Philosophy*, i. Bowling Green, OH: Philosophy Documentation Center.

Doris, J. M. 1996. People Like Us: Morality, Psychology, and the Fragmentation of Character. Ph.D. diss., University of Michigan, Ann Arbor.

Doris, J. M. 1998. Persons, Situations, and Virtue Ethics. *Noûs* 32: 504–30.

Doris, J. M. 2002. *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press.

Doris, J. M., and S. P. Stich. 2001. Ethics. In L. Nadel, gen. ed., D. Chalmers, philosophy ed., *The Encyclopedia of Cognitive Science*. London: Macmillan Reference.

Ellsworth, P. C. 1994. Sense, Culture, and Sensibility. In H. Markus and S. Kitayama, eds., *Emotion and Culture: Empirical Studies in Mutual Influence*. Washington, DC: American Psychological Association.

Firth, R. 1952. Ethical Absolutism and the Ideal Observer Theory. *Philosophy and Phenomenological Research* 12: 317–45.

Flanagan, O. 1991. *Varieties of Moral Personality: Ethics and Psychological Realism*. Cambridge, MA: Harvard University Press.

Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Frankena, W. K. 1976. Obligation and Motivation in Recent Moral Philosophy. In K. E. Goodpaster, ed., *Perspectives on Morality: Essays of William K. Frankena*. Notre Dame, IN: University of Notre Dame Press.

Frankfurt, H. 1988. *The Importance of What We Care About: Philosophical Essays*. Cambridge: Cambridge University Press.

Gibbard, A. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Gilbert, D. T., and P. S. Malone. 1995. The Correspondence Bias. *Psychological Bulletin* 117: 21–38.

Goldman, A. I. 1993. Ethics and Cognitive Science. *Ethics* 103: 337–60.

Haidt, J., S. Koller, and M. Dias. 1993. Affect, Culture, and Morality; or, Is It Wrong to Eat Your Dog? *Journal of Personality and Social Psychology* 65: 613–28.

Haney, C., W. Banks, and P. Zimbardo. 1973. Interpersonal Dynamics of a Simulated Prison. *International Journal of Criminology and Penology* 1: 69–97.

Hare, R. D. 1993. *Without Conscience: The Disturbing World of the Psychopaths Among Us*. New York: Pocket Books.

Hare, R. M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

Harman, G. 1977. *The Nature of Morality*. New York: Oxford University Press.

Harman, G. 1999. Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error. *Proceedings of the Aristotelian Society* 99: 315–31.

(p.281) Harman, G. 2000. The Nonexistence of Character Traits. *Proceedings of the Aristotelian Society* 100: 223–26.

Hart, D., and M. Killen. 1999. Introduction: Perspectives on Morality in Everyday Life. In M. Killen and D. Hart, eds., *Morality in Everyday Life: Developmental Perspectives*. Cambridge: Cambridge University Press.

Hill, T. E. 1991. *Autonomy and Self-Respect*. Cambridge: Cambridge University Press.

Horowitz, T. 1998. Philosophical Intuitions and Psychological Theory. In M. DePaul and W. Ramsey, eds., *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*. Lanham, MD: Rowman and Littlefield.

Hume, D. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. 3rd ed. Oxford: Oxford University Press.

Hume, D. 1978. *A Treatise of Human Nature*. 2nd ed. Oxford: Oxford University Press.

Hursthouse, R. 1999. *On Virtue Ethics*. Oxford: Oxford University Press.

Hutcheson, F. 1738. *An Enquiry into the Original of Our Ideas of Beauty and Virtue, in Two Treatises*. London: D. Midwinter.

Irwin, T. H. 1988. Disunity in the Aristotelian Virtues. *Oxford Studies in Ancient Philosophy*, supp., 61–78.

Isen, A. M., and P. F. Levin. 1972. Effect of Feeling Good on Helping: Cookies and Kindness. *Journal of Personality and Social Psychology* 21: 384–88.

Jackson, F. 1994. Armchair Metaphysics. In J. O'Leary Hawthorne and M. Michael, eds., *Philosophy in Mind*. Dordrecht: Kluwer.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. New York: Oxford University Press.

Jackson, F., and P. Pettit. 1995. Moral Functionalism and Moral Motivation. *Philosophical Quarterly* 45: 20–40.

Johnson, M. 1993. *Moral Imagination: Implications of Cognitive Science for Ethics*. Chicago: University of Chicago Press.

Jones, E. E. 1990. *Interpersonal Perception*. New York: W. H. Freeman.

Kagan, S. 1989. *The Limits of Morality*. Oxford: Oxford University Press.

Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kane, R. 1996. *The Significance of Free Will*. Oxford: Oxford University Press.

Kane, R. 2002a. Introduction: The Contours of Contemporary Free Will Debates. In R. Kane, ed., *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Kane, R. 2002b. Some Neglected Pathways in the Free Will Labyrinth. In R. Kane, ed., *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Keneally, T. 1982. *Schindler's List*. New York: Simon and Schuster.

Kim, J. 1988. What Is “Naturalized Epistemology”? In J. Tomberlin, ed., *Philosophical Perspectives*, vol. 2: *Epistemology*. Atascadero, CA: Ridgeway.

Kitayama, S., and H. R. Markus. 1999. Yin and Yang of the Japanese Self: The Cultural Psychology of Personality Coherence. In D. Cervone and Y. Shoda, eds., *The Coherence of Personality: Social-Cognitive Bases of Consistency, Variability, and Organization*. New York: Guilford Press.

Kupperman, J. J. 2001. The Indispensability of Character. *Philosophy* 76: 239–50.

Larmore, C. E. 1987. *Patterns of Moral Complexity*. Cambridge: Cambridge University Press.

(p.282) Leming, J. S. 1997a. Research and Practice in Character Education: A Historical Perspective. In A. Molnar, ed., *The Construction of Children’s Character: Ninety-Sixth Yearbook of the National Society for the Study of Education*, 11. Chicago: University of Chicago Press.

Leming, J. S. 1997b. Whither Goes Character Education? Objectives, Pedagogy, and Research in Character Education Programs. *Journal of Education* 179: 11–34.

Lewis, D. 1970. How to Define Theoretical Terms. *Journal of Philosophy* 67: 427–46.

Lewis, D. 1972. Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50: 249–58.

Lewis, D. 1989. Dispositional Theories of Value. *Proceedings of the Aristotelian Society*, supp., 63: 113–37.

Loeb, D. 1998. Moral Realism and the Argument from Disagreement. *Philosophical Studies* 90: 281–303.

Louden, R. B. 1984. On Some Vices of Virtue Ethics. *American Philosophical Quarterly* 21: 227–36.

MacIntyre, A. 1984. *After Virtue*. 2nd ed. Notre Dame, IN: University of Notre Dame Press.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. New York: Penguin.

Margolis, E., and S. Laurence. 1999. *Concepts*. Cambridge, MA: MIT Press.

Markus, H. R., and S. Kitayama. 1991. Culture and the Self: Implications for Cognition, Emotion, and Motivation. *Psychological Review* 98: 224–53.

Mathews, K. E., and Cannon, L. K. 1975. Environmental Noise Level as a Determinant of Helping Behavior. *Journal of Personality and Social Psychology* 32: 571–77.

McDowell, J. 1978. Are Moral Requirements Hypothetical Imperatives? *Proceedings of the Aristotelian Society*, supp., 52: 13–29.

McDowell, J. 1979. Virtue and Reason. *Monist* 62: 331–50.

McDowell, J. 1987. *Projection and Truth in Ethics (Lindley Lecture)*. Lawrence: University of Kansas.

McKenna, M. 2001. Source Incompatibilism, Ultimacy, and the Transfer of Non-Responsibility. *American Philosophical Quarterly* 38: 37–51.

Merritt, M. 1999. Virtue Ethics and the Social Psychology of Character. Ph.D. diss., University of California, Berkeley.

Merritt, M. 2000. Virtue Ethics and Situationist Personality Psychology. *Ethical Theory and Moral Practice* 3: 365–83.

Milgram, S. 1974. *Obedience to Authority*. New York: Harper and Row.

Mischel, W. 1968. *Personality and Assessment*. New York: Wiley.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Nagel, T. 1986. *The View from Nowhere*. New York: Oxford University Press.

Nichols, S. 2002. How Psychopaths Threaten Moral Rationalism; or, Is it Irrational to Be Amoral? *Monist* 85: 285–304.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.

Nisbett, R. E. 1998. Essence and Accident. In J. M. Darley and J. Cooper, eds., *Attribution and Social Interaction: The Legacy of Edward E. Jones*. Washington, DC: American Psychological Association.

Nisbett, R. E., and D. Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*. Boulder, CO: Westview Press.

Nisbett, R. E., and L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.

(p.283) Nucci, L. 1986. Children's Conceptions of Morality, Social Conventions and Religious Prescription. In C. Harding, ed., *Moral Dilemmas: Philosophical*

and Psychological Reconsiderations of the Development of Moral Reasoning. Chicago: Precedent Press.

Nussbaum, M. C. 1999. Virtue Ethics: A Misleading Category? *Journal of Ethics* 3: 163–201.

Peterson, D. R. 1968. *The Clinical Study of Social Behavior*. New York: Appleton-Century-Crofts.

Quine, W. V. O. 1969. Epistemology Naturalized. In *Quine, Ontological Relativity and Other Essays*. New York: Columbia University Press.

Railton, P. 1986a. Facts and Values. *Philosophical Topics* 14: 5–31.

Railton, P. 1986b. Moral Realism. *Philosophical Review* 95: 163–207.

Railton, P. 1989. Naturalism and Prescriptivity. *Social Philosophy and Policy* 7: 151–74.

Railton, P. 1995. Made in the Shade: Moral Compatibilism and the Aims of Moral Theory. *Canadian Journal of Philosophy*, supp., 21: 79–106.

Rawls, J. 1951. Outline of a Decision Procedure for Ethics. *Philosophical Review* 60: 167–97.

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

Rosati, C. S. 2000. Brandt's Notion of Therapeutic Agency. *Ethics* 110: 780–811.

Roskies, A. 2003. Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy." *Philosophical Psychology* 16: 51–66.

Ross, L., and R. E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. Philadelphia: Temple University Press.

Saver, J. L., and A. R. Damasio. 1991. Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy Due to Ventromedial Frontal Damage. *Neuropsychologia* 29: 1241–9.

Scanlon, T. M. 1988. The Significance of Choice. In S. M. McMurrin, ed., *The Tanner Lectures on Human Values*, viii. Salt Lake City: University of Utah Press.

Sher, G. 1998. Ethics, Character, and Action. In E. F. Paul, F. D. Miller, and J. Paul, eds., *Virtue and Vice*. Cambridge: Cambridge University Press.

Sherman, N. 1989. *The Fabric of Character: Aristotle's Theory of Virtue*. New York: Oxford University Press.

Shweder, R. A., and E. J. Bourne. 1982. Does the Concept of the Person Vary Cross-Culturally? In A. J. Marsella and G. M. White, eds., *Cultural Conceptions of Mental Health and Therapy*. Boston, MA: Reidel.

Singer, P. 1974. Sidgwick and Reflective Equilibrium. *Monist* 58: 490–517.

Singer, P. 2000. *Writings on an Ethical Life*. New York: HarperCollins.

Sinnott-Armstrong, W. P. 2005. Moral Intuitionism Meets Empirical Psychology. In T. Horgan and M. Timmons, eds., *Metaethics After Moore*. New York: Oxford University Press.

Smart, J. J. C. 1961. Free-Will, Praise and Blame. *Mind* 70: 291–306.

Smith, Adam. 2002. *The Theory of Moral Sentiments*. New York: Cambridge University Press.

Smith, M. 1994. *The Moral Problem*. Oxford: Blackwell.

Stevenson, C. L. 1944. *Ethics and Language*. New Haven: Yale University Press.

Stevenson, C. L. 1963. *Facts and Values*. New Haven: Yale University Press.

Stich, S. 1993a. Naturalizing Epistemology: Quine, Simon and the Prospects for Pragmatism. In C. Hookway and D. Peterson, eds., *Philosophy and Cognitive Science*, Royal Institute of Philosophy, suppl. 34. Cambridge: Cambridge University Press.

Stich, S. 1993b. Moral Philosophy and Mental Representation. In M. Hechter, L. Nadel, and R. E. Michod, eds., *The Origin of Values*. New York: de Gruyter.

(p.284) Strawson, G. 1986. *Freedom and Belief*. Oxford: Oxford University Press.

Strawson, G. 2002. The Bounds of Freedom. In R. Kane, ed., *The Oxford Handbook of Free Will*. New York: Oxford University Press.

Strawson, P. 1982. Freedom and Resentment. In G. Watson, ed., *Free Will*. New York: Oxford University Press.

Sturgeon, N. L. 1988. Moral Explanations. In G. Sayre-McCord, ed., *Essays on Moral Realism*. Ithaca, NY: Cornell University Press.

Sumner, W. G. 1934. *Folkways*. Boston: Ginn.

Svavarsdóttir, S. 1999. Moral Cognitivism and Motivation. *Philosophical Review* 108: 161–219.

Tetlock, P. E. 1999. Review of Culture of Honor: The Psychology of Violence in the South. *Political Psychology* 20: 211–13.

Turiel, E., M. Killen, and C. Helwig. 1987. Morality: Its Structure, Functions, and Vagaries. In J. Kagan and S. Lamb, eds., *The Emergence of Morality in Young Children*. Chicago: University of Chicago Press.

Tversky, A., and D. Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211: 453–63.

Velleman, J. D. 1992. What Happens When Someone Acts? *Mind* 101: 461–81.

Vernon, P. E. 1964. *Personality Assessment: A Critical Survey*. New York: Wiley.

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. 1990. On the Primacy of Character. In Owen Flanagan and Amélie Oksenberg Rorty, eds., *Identity, Character, and Morality: Essays in Moral Psychology*. Cambridge, MA: MIT Press.

Watson, G. 1996. Two Faces of Responsibility. *Philosophical Topics* 24: 227–48.

Weinberg, J., S. Nichols, and S. Stich. 2002. Normativity and Epistemic Intuitions. *Philosophical Topics* 29: 429–60.

Westermarck, E. 1906. *Origin and Development of the Moral Ideas*. 2 vols. New York: Macmillan.

Williams, B. A. O. 1973. A Critique of Utilitarianism. In J. J. C. Smart and B. A. O. Williams, *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Williams, B. A. O. 1981. *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.

Williams, B. A. O. 1985. *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Williams, B. A. O. 1993. *Shame and Necessity*. Berkeley: University of California Press.

Woods, M. 1986. Intuition and Perception in Aristotle’s Ethics. *Oxford Studies in Ancient Philosophy* 4: 145–66.

Woolfolk, R. L., and J. M. Doris. 2002. Rationing Mental Health Care: Parity, Disparity, and Justice. *Bioethics* 16: 469–85

Woolfolk, R. L., J. M. Doris, and J. M. Darley. 2006. Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition* 100: 283–301.

Notes:

(1) . Compare Railton's (1989, 155–56) "methodological naturalism."

(2) . See Gibbard 1990, 58–61; Flanagan 1991; Goldman 1993; Johnson 1993; Stich 1993b; Railton 1995; Blackburn 1998, 36–37; Bok 1996; Doris 1996, 1998, 2002; Becker 1998; Campbell 1999; Harman 1999, 2000; Merritt 1999, 2000; Doris and Stich 2001; Woolfolk and Doris 2002.

(3) . The notation that character is evaluatively independent of or prior to action is something thought to be the distinctive emphasis of virtue ethics (see Louden 1984, 29; Watson 1990, 451–52). But this is not plausibly understood to mean that virtue ethics is indifferent regarding questions of what to do; the question of conduct should be of substantial importance on both virtue and action-centered approaches (see Sher 1998, 15–17).

(4) . Nussbaum (1999, 170) observes that Kantian and utilitarian approaches may share virtue ethics' interest in character. Space prohibits discussion, but if Nussbaum were right, our argument would have more sweeping implications than we contemplate here.

(5) . In Aristotle's view, the virtues are *hexeis* (1984, 1106a10–12), and a *hexis* is a disposition that is "permanent and hard to change" (1984, 8b25–9a9). This feature of Aristotle's account is emphasized by commentators: Sherman (1989, 1) says that for Aristotle (as well as for us) character traits explain why "someone can be *counted on* to act in certain ways" (cf. Woods 1986, 149; Annas 1993, 51; Audi 1995, 451; Cooper 1999, 238).

(6) . This follows quite a standard theme in philosophical writings on virtue and character. For example, Blum (1994, 178–80) understands compassion as a trait of character typified by an altruistic attitude of "strength and duration," which should be "stable and consistent" in prompting beneficent action (cf. Brandt 1970, 27; Dent 1975; McDowell 1979, 331–33; Larmore 1987, 12).

(7) . Part of the reason for this error may be some spirited rhetoric of Harman's (e.g., the title of Harman 2000: "The Nonexistence of Character Traits"). But Harman repeatedly offers qualifications that caution against it; he voices skepticism about the existence of "ordinary character traits of *the sort people think there are*" (1999, 316) and "character traits *as ordinarily conceived*" (2000, 223; our italics). This is to reject a particular conception of character traits, not to deny that character traits exist. For his part, Doris (1998,

507; 2002, 62–66) quite explicitly acknowledges the existence of traits, albeit traits with less generalized effects on behavior than is often supposed.

(8) . The difficulty is not limited to rival academic theories; there is a large body of empirical evidence indicating that everyday "lay" habits of person perception seriously overestimate the impact of individual dispositional differences on behavioral outcomes. For summaries, see Jones 1990; Ross and Nisbett 1991, 119–44; Gilbert and Malone 1995; Doris 2002, 92–106.

(9) . Of course, if the virtue theorist is an elitist, this need not trouble her. But while historical writers on the virtues have at times manifested elitist sympathies (Aristotle 1984, 1123a6–10,1124a17–b32; Hume 1975, 250−67), this is not a sensibility that is typically celebrated by contemporary philosophers.

(10) . There is some question as to whether vices are expected to be robust in the way virtues are, but some philosophers seem to think so: Hill (1991, 130–32) apparently believes that calling someone weak-willed marks characteristic patterns of behavior.

(11) . For example, Kupperman 2001 refers to nine items in the empirical literature in responding to Harman, and Athanassoulis 2000, three.

(12) . A stipulation: We refer to views in the neighborhood of what Darwall (1983, 54) calls "judgement internalism," the thesis that it is "a necessary condition of a genuine instance of a certain sort of judgement that the person making the judgement be disposed to act in a way appropriate to it." Space limitations force us to ignore myriad complications; for more detailed discussion see Svavarsdóttir 1999.

(13) . There is august precedent for supposing that the internalism debate has empirical elements. In his classic discussion, Frankena (1976, 73) observed that progress here requires reference to "the psychology of human motivation"—"The battle, if war there be, cannot be contained; its field is the whole human world." We hope that Frankena would have appreciated our way of joining the fight.

(14) . We prescind from questions as to whether the motivation need be overriding, although we suspect formulations not requiring overridingness are more plausible.

(15) . SCR is a measure of physiological arousal, which is also sometimes called galvanic skin response, or GSR.

(16) . Roskies herself does not offer acquired sociopathy as a counter-example to the Kantian version of empirical internalism, but we believe the evidence *is* in tension with the Kantian view we describe.

(17) . Here Nicholas offers support for the “sentimentalist” tradition, which maintains that emotions (or “sentiments”) play a central role in moral judgment. For a helpful treatment of sentimentalism, see D’Arms and Jacobson 2000.

(18) . These views reflect a venerable tradition linking moral judgment to the affective states that people would have under idealized conditions; it extends back to Hutcheson (1738), Hume (1975, 1978), and Adam Smith (2002).

(19) . A particularly important difference concerns the envisaged link between moral claims and affective reactions. Firth (1952, 317–45) and Lewis (1989) see the link as a matter of meaning, Railton (1986b) as a synthetic identity, and Brandt (1959, 241–70) both as a matter of justification and, more tentatively, as a matter of meaning.

(20) . Brandt was a prolific and self-critical thinker, and the 1959 statement may not represent his mature views, but it well illustrates how empirical issues can impact a familiar approach to ethical theory. For a helpful survey of Brandt’s career, see Rosati 2000.

(21) . On some readings, qualified attitude theories may end up a version of *skepticism* if attitudes don’t converge under ideal circumstances. Suppose a theory holds “an action is morally right (or morally wrong) if all people in ideal conditions would judge that action is morally right (or morally wrong).” Then if convergence fails to obtain in ideal conditions, this theory entails that there are no morally right (or morally wrong) actions.

(22) . This way of putting the argument is at once uncontentious and contentious. It is uncontentious because it does not entail a radical methodological relativism of the sort, say, that insists there is nothing to choose between consulting an astrologer and the method of reflective equilibrium as an approach to moral inquiry (see Brandt 1959, 274–75). But precisely because of this, the empirical conjecture that moral judgments will not converge is highly contentious, since a background of methodological agreement would appear to make it more likely that moral argument could end in substantive moral agreement.

(23) . Strictly speaking, a relativist need not be a “non-factualist” about morality, since, for example, she can take it to be a moral fact that it is right for Hopi children to engage in their fatal play with small animals, and also take it to be a moral fact that it is wrong for American white children to do so. But the factualist-relativist will probably want to reject Smith’s (1994, 13) characterization of moral facts as “facts about the reasons that we all share.”

(24) . See Williams 1985, 136: “In a scientific inquiry there should ideally be convergence on an answer, where the best explanation of the convergence

involves the idea that the answer represents how things are; in the area of the ethical, at least at high level of generality, there is no such coherent hope.”

(25) . The last clause is important, since realists (e.g., Brink 1989, 200) sometimes argue that apparent moral disagreement may result from cultures applying similar moral values to different economic conditions (e.g., differences in attitudes towards the sick and elderly between poor and rich cultures). But this explanation seems of dubious relevance to the described differences between contemporary northerners and southerners, who are plausibly interpreted as applying different values to similar economic conditions.

(26) . The legal scholarship that Nisbett and Cohen (1996, 57–78) review makes it clear that southern legislatures are often willing to enact laws reflecting the culture of honor view of the circumstances under which violence is justified, which suggests there is at least some support among southerners for the idea that honor values should be universalizable.

(27) . In addition to the expedients we have considered, realists may plausibly appeal to, *inter alia*, requirements for internal coherence and the different “levels” of moral thought (theoretical versus popular, abstract versus concrete, general versus particular) at which moral disagreement may or may not be manifested. Brink (1989, 197–210) and Loeb (1998) offer valuable discussions with considerably more detail than we offer here, Brink manifesting realist sympathies and Loeb tending towards anti-realism.

(28) . We think Nisbett and Cohen will fare pretty well under such scrutiny. See Tetlock’s (1999) favorable review.

(29) . There are substantive questions as to what sorts of responses to thought experiments may properly constrain philosophical theory choice. For example, what level of reflection or cognitive elaboration is required: are the responses of interest “pre-theoretical intuitions” or “considered judgments”? We will have something to say about this, but in terminology we will mostly favor the generic “responses,” which we mean to be neutral regarding issues such as cognitive elaboration.

(30) . This analogy with science is not unique to our exposition. Singer (1974, 517; cf. 493) understands Rawls’s (1971) method of reflective equilibrium as “leading us to think of our particular moral judgments as data against which moral theories are to be tested.” As Singer (1974, 493 ff.) notes, in earlier treatments Rawls (1951) made the analogy with scientific theory choice explicit. We needn’t hazard an interpretation of Rawls, but only observe that our analogy is not philosophically eccentric.

(31) . For appeals of this kind, see Blum 1994, 179; G. Strawson 1986, 87–89; P. Strawson 1982, 68; Wallace 1994, 81–82; Williams 1973, 99–100; 1981, 22.

(32) . Some authors—most notably Baron (1994)—have argued that the distorting influences of “heuristics and biases” like those uncovered in the recent psychological literature on reasoning, judgment, and decision-making are widespread in everyday ethical reflection. For overviews of the relevant psychological literature, see Nisbett and Ross 1980; Kahneman, Slovic, and Tversky 1982; Baron 2001.

(33) . We applaud Jackson’s (1998, 36–37) advocacy of “doing serious public opinion polls on people’s responses to various cases.” However, we expect this may be necessary more often than Jackson imagines. According to Jackson (1998, 37), “Everyone who presents the Gettier cases [which are well-known epistemology thought experiments] to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases.” Yet Weinberg et al. (Weinberg, Nichols, and Stich 2002) found that responses to epistemology thought experiments like the Gettier cases varied with culture and SES; this suggests that philosophers need to be more systematic in their fieldwork.

(34) . Our formulation is meant to be quite standard. Kane (2002a, 10) observes that statements of the difficulty in terms of alternative possibilities have dominated modern discussion. A recently prominent formulation proceeds not in terms of PAP, but by way of an “ultimacy condition,” which holds that an actor is responsible for her behavior only if she is its “ultimate source” (see McKenna 2001, esp. 40–41). This does not impact the present discussion, however. First, notice that although some may maintain an ultimacy requirement and reject PAP, the two commitments need not be incompatible; Kane (1996, 2002b) holds them both. Secondly, as should become clear, the empirical work we describe below is relevant to both formulations.

(35) . As Kane (2002a, 27–32) observes, relatively few philosophers have been unqualifiedly committed to hard determinism; Smart’s (1961) views on responsibility, for example, are complex.

(36) . Peter Strawson (1982) did the pioneering philosophical work on the reactive attitudes; he appears to reject the suggestion that such attitudes manifest a commitment to something in the spirit of incompatibilism.

(37) . G. Strawson puts the point rather emphatically, but similar observations are commonplace. Cf. Nagel 1986, 113, 125; Kane 1996, 83–85.

(38) . There is again a question about the scope of “people”; Strawson’s reference to “natural facts” may suggest that he is making a boldly pancultural attribution, but he might be more modestly attributing the theory only to those people who embody something like the “Western ethical tradition.” We will not

attempt to decide the interpretative question, because the empirical work we describe troubles even the more modest claim.

(39) . G. Strawson (1986, 25–31; 2002) may favor formulations in terms of ultimacy rather than PAP (see n. 33 above). This doesn't affect our argument, since the empirical work we recount below looks to trouble a prediction formulated in terms of ultimacy as well as the alternative possibilities formulation we favor.

(40) . For some discussion, see Velleman 1992; Bratman 1996; Watson 1996; Doris 2002, 140–46.

(41) . Woolfolk et al. (2006) obtained similar results for the prosocial behavior of kidney donation: an identified actor was credited for making a donation even when heavily constrained.

(42) . It also looks as though the actor fails an ultimacy condition (see nn. 34 and 39 above).

(43) . For much valuable feedback, we are grateful to audiences at the Moral Psychology Symposium at the 2001 Society for Philosophy and Psychology meetings, the Empirical Perspectives on Ethics Symposium at the 2001 American Philosophical Association Pacific Division meetings, and a series of lectures on philosophy and cognitive science held at the Australian National University in July 2002—especially Louise Antony, Daniel Cohen, Frank Jackson, Michael Smith, and Valerie Tiberius. Thanks to Daniel Guevara, Jerry Neu, Alva Noë, and especially Don Loeb, Shaun Nichols, and Adina Roskies for comments on earlier drafts.