Can Innate, Modular "Foundations" Explain Morality? Challenges for Haidt's Moral Foundations Theory

Christopher L. Suhler and Patricia Churchland

Abstract

■ Jonathan Haidt's Moral Foundations Theory is an influential scientific account of morality incorporating psychological, developmental, and evolutionary perspectives. The theory proposes that morality is built upon five innate "foundations," each of which is believed to have been selected for during human evolution and, subsequently, tuned-up by learning during development. We argue here that although some general elements of Haidt's theory are plausible, many other important aspects of his account are seriously flawed. First, *innateness* and *modularity* figure centrally in Haidt's account, but terminological and conceptual problems foster confusion and ambiguities. Second, both the theory's proposed number of moral foundations and its taxonomy of the moral domain appear contrived, ignoring equally good candidate foundations and the possibility of substantial intergroup differences in the foundations' contents. Third, the mechanisms (viz., modules) and categorical distinctions (viz., between foundations) proposed by the theory are not consilient with discoveries in contemporary neuroscience concerning the organization, functioning, and development of the brain. In light of these difficulties, we suggest that Haidt's theory is inadequate as a scientific account of morality. Nevertheless, the theory's weaknesses are instructive, and hence, criticism may be useful to psychologists, neuroscientists, and philosophers attempting to advance theories of morality, as well as to researchers wishing to invoke concepts such as innateness and modularity more generally. ■

INTRODUCTION

Morality permeates human existence, playing a role in a vast array of choices and evaluations, both public and private, momentous and trifling. Given morality's central place in human social life, it is important for scientists to address its origin and features. Although the scientific study of morality is still in its infancy, research is underway on topics including the role of emotion in moral judgment (e.g., Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001), the neural bases of punitive behavior (e.g., de Quervain et al., 2004) and cooperative behavior (e.g., Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Rilling et al., 2002), and the evolutionary roots of morality (e.g., Bowles, 2008; Casebeer, 2003; Gintis, Bowles, Boyd, & Fehr, 2003).[1]

Still needed, however, is a comprehensive explanatory framework within which to situate these findings and numerous others concerning phenomena such as moral development and inter- and intragroup moral differences. Does such a unifying framework currently exist? Jonathan Haidt thinks so. He, along with collaborators including Craig Joseph, Jesse Graham, and Brian Nosek, has proposed a theory, the Moral Foundations Theory (MFT), according to which morality rests upon five innate "foundations" (Haidt & Joseph, 2007).[2] MFT's influence extends beyond academia, having received favorable coverage in popular media outlets (e.g., Brooks, 2009; Shermer, 2009; Wade, 2007).

Although some aspects of MFT are likely broadly correct, our analysis shows that most of its defining features are seriously flawed. First, the terminological and conceptual apparatus on which MFT relies—particularly as it relates to *innateness* and *modularity*—is rife with confusion and ambiguities that leave it either highly implausible or explanatorily toothless (see Terminological & Conceptual Confusions section). Second, the theory's taxonomy of the moral domain appears contrived, sidelining equally good candidate foundations and neglecting substantial intergroup differences in the foundations' contents (see Difficulties with the Five-Foundations Taxonomy section). Third, the mechanisms (viz. informationally encapsulated, domain specific modules) appealed to by MFT do not comport well with discoveries in contemporary neuroscience concerning the organization, functioning, and development of the brain (see Neurobiology and MFT section). In light of these difficulties, we suggest that Haidt's account is currently inadequate as a scientific theory of morality.

AN OVERVIEW OF MFT

MFT's essential idea is that morality is made up of five[3] sets of foundations, each set corresponding to an adaptive challenge faced by early humans and their nonhuman primate

University of California, San Diego

ancestors. The five foundations, followed by their respective adaptive challenges, are:

1. *harm/care*—protect and care for young, vulnerable, or injured kin
2. *fairness/reciprocity*—reap benefits of dyadic cooperation with nonkin
3. *ingroup/loyalty*—reap benefits of group cooperation
4. *authority/respect*—negotiate hierarchy, defer selectively
5. *purity/sanctity*—avoid microbes and parasites[4]

Given the importance of these five domains to a group-living species such as *Homo sapiens*, Haidt suggests that evolution likely would have built them into us in some manner; they would have become *innate* (Haidt & Joseph, 2007). By invoking innateness, however, Haidt is not claiming that the foundations are, in their full-dress form, biologically "hardwired," as is, say, the eye blink reflex. Rather, the notion of innateness in play is one of *preparedness*; following Marcus (2004), Haidt takes the foundations to be innate in the sense of being "organized in advance of experience" (Haidt & Graham, 2009, p. 382, quoting Marcus, 2004, p. 40).

In what sense, specifically, might humans be "prepared" to acquire and deploy moralities corresponding to the five foundations? Here, Haidt invokes the notion of mental modularity, originally articulated by Fodor (1983). In line with its conception of innateness, MFT's modules are taken to be dependent on environmental, and not just genetic, factors. Moreover, the definition of modularity is a loosely inclusive one; drawing on the work of Sperber (1994, 2005), Haidt says that a putative module needs only to be "modular to some interesting degree" (Haidt & Joseph, 2007, p. 380, quoting Sperber, 1994[5]) to be counted as such.

More specifically, the modules that "prepare" humans to acquire certain moral concerns are what Haidt, again following Sperber (2005), calls "learning modules." These correspond to each of the five foundations and facilitate the learning of specific responses to patterns of events in the social world.[6] Continuing his utilization of Sperber's "teeming modularity" framework, Haidt contends that the outputs of these learning modules are, themselves, a plethora of modules, each of which is dedicated to producing some situation-specific moral intuition falling within one of the five foundations (Haidt & Joseph, 2007). The modules so produced are "like little bits of input–output programming connecting the perception of a pattern in the social world (often a virtue or vice) to an evaluation and in many cases a specific moral emotion (e.g., anger, contempt, admiration)" (Haidt & Joseph, 2007, pp. 379–380).[7] (We will henceforth call these *second-order modules* as a convenient way of distinguishing them from learning modules.)

Second-order modules produce moral intuitions—flashes of approval, disapproval, or other emotions upon detecting some pattern in the social world (Haidt & Joseph, 2007, p. 379). These intuitions, in turn, are what actually drive our moral judgments, beliefs, actions, and the like. This is an important point of divergence between MFT and views of morality historically accepted in philosophy (e.g., Kant, 1785/1998) and psychology (e.g., Kohlberg, 1981, 1984), which have tended to focus on reason as the sole driver of moral judgment.[8] On Haidt's view, by contrast, reasoning plays a mostly post hoc role, coming into play only after (affective) intuitions produce a moral judgment and serving to justify that judgment to others (see also Haidt, 2001).

A second way in which MFT differs from many other accounts of morality is in the breadth of the moral domain that it posits. Certain other theories (e.g., Turiel, 1983) seem to restrict morality to the harm/care and fairness/reciprocity foundations. This, Haidt notes, accounts reasonably well for the moralities of "modern" (i.e., western, secular, liberal) cultures. However, Haidt and Graham (2009) argue, on the basis of cross-cultural data (e.g., Shweder, Much, Mahapatra, & Park, 1997; Haidt, Koller, & Dias, 1993), that these theories give a very incomplete picture of morality as manifested in "traditional" cultures—a category which includes many nonwestern cultures as well as western religious conservatives. The moralities of these groups do include concern for the first two foundations, but they also include a high degree of concern for the other three—ingroup/loyalty, authority/respect, and purity/sanctity.

This approach, Haidt suggests, provides an elegant developmental explanation of moral difference: Although all people are "prepared" to develop concern for all five foundations (i.e., they are genetically endowed with learning modules for all five foundations), the environment in which they are raised may result in second-order modules developing for only some of these domains (at least to any substantial degree). Thus, a person raised by liberal parents may develop second-order modules concerned principally with harm/care and fairness/reciprocity, whereas her counterpart raised by conservative parents may develop second-order modules concerned with all five foundations to a similar (high) degree. In this way, MFT purports to be able to explain both moral similarity and difference between individuals and cultures.

We have aimed to capture the essence of Haidt's theory as best we can, although of course, many details have been glossed over for reasons of brevity. Despite our criticisms of many pillars of MFT in what follows, we wish to pause briefly to acknowledge a number of its genuine strengths. One strength is MFT's recognition of emotion and intuition as important contributors to moral behavior and judgment, an aspect of the theory that is consistent with a long tradition stretching from the 18th-century philosophy of David Hume (1739/1740/2008; for discussion, see Blackburn, 2008) to present-day cognitive science (see note 7). Another is the theory's rejection of the common philosophical (and, sometimes, psychological) project of searching for a single, unified principle or psychological mechanism that, by itself, accounts for all of morality, a project that has thus far proved something of a quagmire.[9] A further strength is MFT's emphasis on the need to examine cross-cultural

data to see whether the moral domain has a different composition outside of the western cultures in which most scientific research on morality has been conducted. Finally, we commend Haidt's recognition of the need to provide a developmental and evolutionary account of morality, a project that is too often overlooked by researchers focusing on the proximate mechanisms of morality as manifested in adults.[10]

Despite these strengths, we believe that crucial aspects of MFT are implausible or, at a minimum, in need of substantial rethinking. These difficulties with the theory are the subject of the remainder of this article.

TERMINOLOGICAL AND CONCEPTUAL CONFUSIONS

Innateness

Innateness is a notoriously treacherous concept that is used in many different ways by many different researchers.[11] Recent advances in embryology and molecular biology have rendered dubious any neat division of traits into those that are innate and those that are acquired through experience by making it clear that behavior, in general, is the outcome of interactions between genes, brain, and behavior/experience (Flint, Greenspan, & Kendler, 2010; Richerson & Boyd, 2005; Greenspan, 2001). Even so, as Richerson and Boyd (2005, p. 9) point out, one can reject the simplistic nature–nurture dichotomy but still acknowledge that "different traits do vary in how sensitive they are to environmental influences." Hence, to avoid mere hand-waving, innateness claims have to provide evidence that the traits they target tend to the "insensitive-to-environmental-influences" end of the spectrum, *and*, for adaptationist accounts, that these traits were selected for in the course of human evolution. Thus, they have to tread a fine line: On the one hand, too strong a notion of innateness is likely not to be applicable to cognitive and behavioral traits of any complexity, morality included, because so much learning is involved. On the other, too weak a notion may apply to far too *many* cognitive and behavioral traits to be useful for explanation, prediction, and categorization.

Although Haidt generally seems to recognize the traps in the unduly strong version, he is less sensitive to those at the weak end of the spectrum. The result is that his version of nativism excludes very little, being so loose that it applies to nearly any trait displayed in human behavior. This promiscuous applicability compromises its power to categorize traits or to generate detailed predictions about their evolutionary and developmental origins, or cognitive mechanisms. One of the few clear predictions generated by the theory is that some moral norms will be more easily learned than others; this prediction is central to Haidt's definition of innateness as "preparation for" or "organiz[ation] in advance of experience." Such definitions, however, risk triviality due to the absence of details concerning what, precisely, this evolved preparedness or pre-experiential organization amounts to. Some things most humans can easily learn are not plausibly innate in any meaningful sense, such as how to ride a bicycle, tie a reef knot, use a spoon, paddle a boat, or use a phone. Nor is the ease or speed of acquisition always a reliable marker of innateness in Haidt's sense. Children, for example, can frequently learn to open drawers and untie shoelaces before they can walk, something the neurobiological evidence shows they *are* neurally prepared to do.[12]

This problem could, to some extent, be avoided if Haidt supplied some specifics of how—at the level of cognitive psychology, developmental psychology, neuroscience, and so on—humans are "prepared for" the acquisition of some moral norms but not others. Unfortunately, he provides nothing of the sort. Consequently, dubbing a given trait "prepared for" amounts to little more than a restatement of the behavioral data, providing little basis for nontrivial empirical evaluation of the claim.

Here, Haidt might reasonably protest that he does provide a more specific account of the sort of preparedness at work: learning modules. If there is a learning module for traits or abilities falling within some domain, these traits or abilities are innate (in the sense of being organized/prepared for). Without such a module, the trait or ability is not innate. But how do we know when a trait emerges from a learning module and when it does not?

Modularity

As with *innateness*, MFT's invocation of the concept of modularity is murky, raising the question of whether it is truly substantive, or merely conveniently vague and context-morphable. Had Haidt made a strong modularity claim along the lines of those made by certain evolutionary psychologists (e.g., Cosmides & Tooby, 1994, 1997), it would at least qualify as substantive (in some minimal sense). Regrettably, however, it also would almost certainly be incorrect, as such claims are neurobiologically very dubious, however tempting they may seem at a psychological level.[13] Alternatively, if a weaker notion of modularity is used, there arises the question of whether the notion actually explains the phenomenon in some deeper, more informative way, or whether it is merely a vague, "black-box" restatement of the behavioral data, lacking computational, neurobiological, or other details. This latter difficulty is the principal one afflicting MFT.

What, precisely, are the criteria Haidt associates with modularity? So far, as we can tell, there are two. The first is *domain specificity*. In the case of learning modules, the domains in question are the five foundations. The modules respond to and facilitate learning of moral norms falling within these foundational domains. In the case of second-order modules, they are particular patterns in the social world (Haidt & Joseph, 2007, pp. 379–380), such as seeing a fellow shopper cut in line or a hunter club a seal. The second criterion is *informational encapsulation*. This

criterion, unlike the first, is not made entirely explicit in some of Haidt's published work, but Haidt (personal communication, May 5, 2010) explains that it was, in fact, what first drew him to modularity theories. The link was his earlier work on the phenomenon of "moral dumbfounding," in which individuals cling to their intuitive moral judgments despite being unable adequately to justify them. The individuals, moreover, may continue to endorse their initial judgments even when shown that the justification they attempted to offer (e.g., saying that there is some harm involved) is explicitly ruled out by the terms of the scenario they are judging (see Haidt, 2001; Haidt et al., 1993). For Haidt, this perseveration of certain moral judgments, despite new or confounding information, implied that the processes producing the judgments are informationally encapsulated (to some degree) and, therefore, that the judgments are the product of mental modules.

We suggest that these criteria are inadequate to ground an informative, explanatorily useful notion of modularity. Beginning with Haidt's version of domain specificity, the problem, as we analyze it, is that the modules of MFT appear to be domain-specific only in the vacuous sense that whatever set of input–output (i.e., stimulus–behavior) patterns happens to be manifested is said to be the result of modular processing and to constitute the domain to which that module is dedicated. This, of course, robs the view of any distinguishing power because *any* cognitive process or trait qualifies as domain-specific on this trivial description. This is particularly—but not solely—a problem for second-order modules, which seem to be little more than a way of designating somewhat arbitrarily chosen (and, perhaps, arbitrarily fine-grained) stimulus–behavior patterns without shedding any light on the underlying processes' computational workings or other features. Second-order modules are thus examples of "black-box" psychology *par excellence*, giving the illusion of explanation and understanding but, in fact, providing very little of either. Conceivably, neurobiological data might shore up the idea, but as we show below (Neurobiology and MFT section), this seems unlikely.

Learning modules face similar difficulties. Although Haidt is undoubtedly correct that certain sorts of moral norms are more easily learned than others, the inference to domain-specific learning modules for these easily learned norms is completely unsupported by the available evidence. The mere commonness of moral norms corresponding to the five foundations and the ease with which some adaptationist account can be concocted for this commonness do not sanction the move to something as concrete as domain-specific, genetically endowed learning modules.[14] Myriad other psychological explanations, with far greater convergent support from neurobiology and other fields, could account for the phenomena. Broadly speaking, these explanations involve the combined effect of numerous nondedicated processes and their interactions, a possibility we examine in greater detail in the Neurobiology and Domain Specificity section below.

How does informational encapsulation fare? Lack of evidence and lurking triviality are again the principal difficulties. Triviality comes in when Haidt declines the clear-cut type of informational encapsulation of other theorists (e.g., Fodor, 1983), preferring instead "partial" or "to-some-degree" encapsulation. Unfortunately, nearly *any* higher cognitive process can lay as much claim to such "partial encapsulation" as moral dumbfounding, the phenomenon Haidt takes to be paradigmatic of encapsulation in the moral domain. Psychological research—to say nothing of everyday experience—provides a host of examples of situations in which people are insensitive to or distort novel or potentially threatening information.[15] A cuckolded husband may maintain the comforting belief that his wife is faithful even in the face of overwhelming evidence to the contrary. A loving mother may refuse to believe her wonderful son misbehaves in school. An experimental subject may persist in believing that she is adept at a certain task even after being told that the feedback on which she based this belief was entirely bogus (e.g., Ross, Lepper, & Hubbard, 1975). And, to take an example involving science, a researcher may cling to a preferred theory even after powerful arguments and evidence have accumulated against it. For example, Joseph Priestley, the brilliant English chemist, espoused to his death the theory that combustion and rusting involved release of phlogiston, rejecting the increasingly well-confirmed oxidation theory of Lavoisier. More recently, Fred Hoyle, despite mounting evidence against him, clung to the steady-state theory of the origins of the universe. Following Haidt, one might call this "scientific dumbfounding," and propose an encapsulated module to explain it.

Perseveration in these cases does not, of course, necessarily mean that the beliefs are the product of modules, for there are probably motivational, affective, and other nonmodular factors that are more than sufficient to account for the phenomenon. Ironically for Haidt, an explanation invoking such factors actually counts against encapsulation, insofar as it suggests robust links between affective processes, on the one hand, and those involving factual beliefs, on the other.[16]

Setting aside the above difficulties concerning innateness and modularity, might Haidt's five foundations still be a productive way to understand morality? In the next section, we examine the five-foundations taxonomy and its rationale.

DIFFICULTIES WITH THE FIVE-FOUNDATIONS TAXONOMY

Additional Foundations

One shortcoming of the five-foundations taxonomy is that other basic moral values exhibited by humans across various cultures have as much—or as little—call to be included as do Haidt's favored five.[17] Two strong contenders for this role are *industry* and *modesty*.[18]

Industry is highly moralized in many societies—including modern, secular western ones, and groups as varied as the Inuit, Scots, Japanese, Haida, Finns, and Dene, to name a few. It is manifested in a strong "work ethic," a repudiation of laziness, and a disapproval of various other forms of shirking and free-riding. This value on industry seems to exist above and beyond any potentially harmful effects sloth might have on individual, community, or purity/sanctity considerations, and on top of any connections hard work may have to hierarchical roles (e.g., working because a superior directs one to, providing for lower-status individuals). Fairness/reciprocity seems to be the only current foundation into which one might try to slot industry. This, however, fails given that hard work seems to be valued and laziness scorned even when a lack of industry would not compromise one's ability to meet obligations incurred in the course of cooperative or reciprocal interactions with other individuals.

As for modesty, many societies have subtle norms proscribing overtly calling attention to one's achievements, status, wealth, and so forth. Bragging behavior is frowned upon except in a limited range of situations. For example, in western societies such as the United States, a child may boast to her parents about some achievement in school or on a sports team, but doing the same around her friends would likely result in her being shunned and labeled a braggart. As with industry, finding a niche for modesty in the five current foundations is difficult. Authority/respect is the only one into which modesty might plausibly be slotted, but at present, this foundation deals rather specifically with behaviors related to hierarchy. Although behaviors related to modesty/immodesty could be described, at least in part, in terms of their role in helping one navigate the social hierarchy, modesty seems, like industry, to be valued even in situations where such hierarchical concerns are not present.

Content of the Foundations

The converse of omitting some basic moral values is plugging in too many, and here too, Haidt's taxonomy looks worryingly ad hoc rather than principled. For example, the ingroup/loyalty and purity/sanctity foundations may, for all we can be sure, merely be extensions of harm concerns to entities other than individual persons—for instance, to a supraindividual entity, the community, the welfare of which is taken to be more than the mere sum of its individual members' welfares. Purity/sanctity, likewise, may concern perceived harms to supernatural entities not acknowledged by secular morality: deities, souls, and the like. Violations of purity/sanctity norms may also take the form of using something—one's body or a sacred site, for example—in a way that does not accord with the intentions/wishes of deities, thereby harming and defiling it. Placating the gods or spirits or what have you to prevent natural disasters and diseases, and blaming the failure to do this when these or other calamities occur, is entirely common, suggesting a further way in which purity/sanctity may be linked to harm.

Also infected by the ad hoc problem is the list of "characteristic emotions" paired with each foundation. For example, Haidt lists anger as a characteristic emotion of the fairness/reciprocity foundation (see, e.g., Haidt & Joseph, 2007, p. 382, Table 19.1). Yet anger can just as naturally be deployed in the service of many of the other foundations, and perhaps all of them. Most obviously, it is a crucial and ubiquitous aspect of defensive responses of the sort associated with the harm/care foundation, although Haidt does not list it as such. For example, females of many mammalian species, humans included, become impressively hostile toward predators and conspecifics who threaten their young. Anger is also routinely triggered by events associated with the other putative foundations, such as a failure of conspecifics to show proper deference or being "dissed" by a peer or subordinate (authority/respect), malicious actions that undermine the group (ingroup/loyalty),[19] or the flouting of religious norms (purity/sanctity), not to mention various allegedly nonfoundational violations, such as bragging, lying, and shirking. Other "characteristic" emotions, such as guilt and disgust, likewise seem to have substantially broader moral expression than their "characteristic" association with particular foundations claimed by MFT would imply.

These criticisms point to yet another murky issue: What counts as an instance of a behavior falling within a given foundational category? This difficulty is particularly acute for the purity/sanctity foundation. Even within the rather coarse category of "traditional" societies, there is huge variability in religious practices, with many religions not invoking a Deity or a set of gods as such, but rather assorted spirits, ancestral ghosts, or merely beliefs in the efficacy of sacrifices and other rituals. This diversity of content is important because it bears on Haidt's claim that a signal advantage of MFT is that it allows one to understand differences in the moral profiles of "traditional" and "modern" groups in terms of the differential development of the five foundations. Fair enough, MFT does, at first, appear to provide a natural explanation for such differences (see Haidt & Graham, 2009; Haidt & Joseph, 2004, 2007) and to have experimental evidence supporting this contention (see, e.g., Graham, Haidt, & Nosek, 2009). A closer look suggests, however, that this apparent achievement for MFT is, in fact, no such thing, and its seeming otherwise is simply due to the fact that Haidt has gerrymandered the content of the three "traditional" foundations (ingroup/loyalty, authority/respect, and purity/sanctity) in such a way as to make this outcome inevitable. Our contention is this: By and large, liberals also have significant moral concerns that arguably fall within these three categories; the *contents* or *targets* of these concerns just happen not to be the same as those of conservatives.

As an illustration, consider again the purity/sanctity foundation. Haidt argues that the adaptive challenge underlying

this foundation—and its "characteristic" emotion, disgust—is the need to avoid microbes and parasites in one's food. This system for evaluating and rejecting potentially illness-causing food, he suggests, was then exapted for "social evaluation and rejection" (Haidt & Joseph, 2007, p. 384). Although one may quibble with the details, this is a reasonable enough hypothesis about the origin of disgust's role in morality. However, when Haidt specifies the sort of contemporary moral concerns associated with the purity/sanctity system, he focuses on the "sanctity" aspect of the putative foundation, very narrowly interpreted. Consequently, his examples, and the resulting conception of the foundation, end up being confined to those that some *religious* individuals consider important, such as "keeping religious objects set apart from pollutants and profane objects [and] overcoming carnal desires and treating the body as a temple... [as well as] virtues such as chastity and temperance, and vices such as lust and intemperance" (Haidt & Joseph, 2007, p. 384).

This neglects the possibility that liberals also make substantial use of the purity/sanctity system. For example, at present, many liberals are extremely concerned about the environment. These concerns are often highly moralized, with members of groups such as Greenpeace and Earth First being every bit as sanctimonious about damage to old growth forests, the clearing of tropical rainforests for palm plantations, and the pollution of watersheds as conservatives allegedly are about issues such as promiscuity, homosexuality, and the defilement of religious objects. Their concerns about these issues may be voiced through use of the language of purity, defilement, holiness, and sanctity that is every bit as sincere as that of religious conservatives. To claim that their behavior is not "true" purity/sanctity behavior simply because they do not believe in a Personal Deity would require some fancy, and almost certainly ad hoc, maneuvering.

Similar points can be made about liberals who care deeply about "animal rights." They may refuse to wear fur or leather, not consume meat or even any animal products at all, or eat only organic food—and feel disgust toward people and companies that do consume or produce these things. Activists may also accuse those who do not share their views of wanton immorality through public protests and campaigns. Just as certain conservative groups may rally at a state capitol to protest the legalization of same-sex marriage, members of People for the Ethical Treatment of Animals may gather outside a dog show dressed in Ku Klux Klan regalia to decry the show's supposed efforts to create a "master race" of purebred dogs (see Farris, 2009).

In light of this, it seems likely that if subjects in studies such as that of Graham et al. (2009) were given questions pertaining to a broader range of purity/sanctity concerns—including not only things like chastity and sin but also respecting the environment and animal rights—the gap between conservatives' and liberals' concern with *this* foundation might well disappear entirely.[20]

Gradations between "Modern" and "Traditional"

Haidt often touts his theory's seeming ability to capture a contrast between the moralities of "modern" (or "liberal") and "traditional" (or "conservative") groups. He claims that the former's morality consists principally of the harm/care and fairness/reciprocity foundations, whereas the latter's involves substantial development of these two foundations plus the three others, ingroup/loyalty, authority/respect, and purity/sanctity. This supposed discovery has been excitedly picked up on by the popular press (e.g., Shermer, 2009).

We are more skeptical, however, as examples abound of individuals and groups that do not fit neatly into the categories of "modern" or "traditional."[21] Many twentieth- and twenty-first-century Swedes, Danes, Germans, and French, for example, have not only the expected "modern" moral concerns but also concerns for social welfare and collective/national identity that the "modern" foundations alone seem insufficient to account for. And what of individuals who, in the United States, identify themselves as "libertarians" or as "socially liberal and economically conservative"? For libertarians, harm/care and fairness/reciprocity values are evident, but given the differences between the moral judgments, voting patterns, and so forth of these individuals and American liberals—whom Haidt also takes to have predominantly harm/care and fairness/reciprocity concerns—there must be other factors that set the two groups' moralities apart.

Can MFT account for the existence of these groups in a systematic, evidentially powerful way? First appearances are promising, as MFT's approach of positing separate learning modules for each of the five foundations seems to provide a convenient way of accounting for the possibility of such independent development.[22]

We suggest, however, that the moral values of groups that are not paradigmatically "modern" or "traditional" are not so smoothly explained by MFT. It seems a stretch, for instance, to reduce a western European's concern for social welfare to the development of the ingroup/loyalty foundation in the same way that Haidt (e.g., Haidt & Joseph, 2007, p. 383; Graham et al., 2009) sometimes suggests that an American conservative's nationalism might be so reduced. Indeed, after the horrors of two world wars, the nations of western Europe are extremely wary of anything that smacks of nationalism or militarism. Similarly, if one accepts Haidt's contention that American liberals have only very limited development of the three "traditional" foundations, it is not clear how their foundational repertoire differs from that of libertarians. Nor are the differences between these groups explicable in terms of the relative development of the two "modern" foundations; both seem to have very high concern for both harm/care and fairness/reciprocity.

These difficulties call attention to a methodological flaw hitherto unremarked in our discussion: the limitations of the dataset Haidt has relied upon in constructing his theory. He has, in effect, selected two points on a broad and varied

moral landscape—"modern"/"liberal" and "traditional"/"conservative" morality—and tailored his theory to account for them. If he had selected two (or more) different points on which to base his theory—for instance, western Europeans and American libertarians—the set of foundations that he came up with might have been quite different. At a minimum, the contents associated with each foundation may have differed substantially.[23] Given this, MFT may be open to the same charge of myopia that Haidt has leveled at other researchers' views (see, e.g., Haidt & Graham, 2009; Haidt & Joseph, 2007, Section 2).

NEUROBIOLOGY AND MFT

Why Neurobiology Matters

Progress in neurobiology, developmental psychology, and genetics in recent decades means that innateness hypotheses are now expected to be supported by, or at least consilient with, evidence from these fields. Long gone are the days when conjectures about universality and selective advantages would suffice. Supporting evidence is manifestly required for strong forms of nativism that take complex cognitive traits to be completely genetically specified, requiring little environmental input for their proper development. However, evidence of this sort is also needed for weaker forms of nativism such as Haidt's "preparedness," all the more because such data could firm up awkward vagueness in behavioral criteria.

Our point is not that anyone utilizing nativist concepts in their accounts of morality must also commit themselves to an *identification* of moral kinds with neurobiological kinds. Like Haidt, we are skeptical that any such neat identity relationship will be found. But acknowledging this inevitable messiness does *not* absolve researchers invoking innateness from the responsibility of having a theory that is consilient with what is known in neurobiology. To belabor the obvious, if a theory posits certain phenomena as owed to evolution by natural selection, this implies that the phenomena have at least *some* basis in our evolved biology. And in the case of cognition, behavior, and learning, the primary biological locus of these functions is the brain.[24] Thus, if there are Haidt-type mental modules, we should expect the organization of the brain to reflect or, at an absolute minimum, be consistent with such modularity.

Neurobiology and Informational Encapsulation

Let us start with informational encapsulation. In the mammalian brain, the neuroanatomical rule is "loopy" (reentrant, recurrent, etc.) architecture, not informationally encapsulated, feedforward modules (Logothetis, 2008). The cortex also has a "small world" architecture, meaning that local connections are dense and long-range connections are sparse, but everything is easily accessible to everything else in a few synaptic steps (Bullmore & Sporns, 2009; Buzsáki, 2006). This pattern is typical even in primary visual cortex (V1), the first area of cortex to receive inputs from the retina (via the lateral geniculate nucleus, or LGN), where more than 80% of the synaptic contacts on V1 neurons come *not* from the LGN, but from elsewhere, in particular, from higher levels in the visual system, other regions of the thalamus, and so forth. Additionally, every area of the cortex is "loopily" connected to the thalamus (Sherman, 2005). One region of the thalamus—the intralaminar nuclei—projects reciprocally to every part of the cortex, save V1. These data call into question the neuroanatomical plausibility of informational encapsulation.

Physiologically, informational encapsulation is undermined by the fact that cortical tissue regularly, not rarely, exhibits spontaneous activity, that is, measured neuronal spiking not occasioned by external stimuli or attention to a task. As Ringach (2009) points out, experiments using multielectrode techniques, as well as those using voltage-sensitive dyes, have shown that spontaneous activity of single cells is structured in space (coherent across millimeters of cortical tissue) and at various time scales. This activity can be related to hallucinations during sensory deprivation tests and to subsequent task performance. Super, van der Togt, Spekreijse, and Lamme (2003) showed that higher levels of *spontaneous* activity in V1 prior to presentation of a difficult task increased the monkey's successful performance of the task, a result which challenges the traditional hunch that spontaneous activity of single cells is just "noise." Imaging techniques have revealed increases in the level of activity seen in the so-called default network during the nontask state (e.g., Buckner, Andrews-Hanna, & Schacter, 2008; Gusnard & Raichle, 2001). The fact that nontrivial amounts of energy are used by spontaneous activity of this sort provides further reason to think that it is not just neuronal "idling" or noise.

Even those processes in so-called early perception, advertised by Fodor (1983) as the "obvious" best candidates for modularity, fail to satisfy the requirement of informational encapsulation. Cells as early in visual processing as the LGN—not even yet in cortex—are sensitive not only to visual input but also to task demands and motor planning (Casagrande, Sáry, Royal, & Ruiz, 2005). This is owed to input from other regions, including the brainstem, that feeds into the LGN. At the single-cell level, Paradiso, MacEvoy, Huang, and Blau (2005) showed that there is extra-receptive field modulatory input to cells in V1. Niell and Stryker (2010), in a landmark study, demonstrated that single neurons in V1 double their firing rate in response to the *very* same stimulus depending on whether the motor system is engaged or not (i.e., on whether the animal is walking in a wheel). A similar result was found in *Drosophila*: Neural responses to precisely the same visual stimulus vary as a function of whether the fly is flying or not (Maimon, Straw, & Dickinson, 2010). These data challenge the idea that cells processing early visual information are shielded from activity in distant parts of the brain, suggesting instead that the processes of "early"

perception are quite informationally porous, albeit in ways not clearly understood.[25]

Experiments on humans using neuroimaging show that visual processing areas (including V1) can be recruited for purposes of decoding language (Sadato et al., 1996), as well as for problem-solving tasks and when engaging in mental simulation (Kosslyn & Thompson, 2003; Kosslyn, Ganis, & Thompson, 2001; Mellet et al., 2000), violating not only the requirement of informational encapsulation but also that of domain specificity. Recent brain imaging studies in blind subjects have provided additional support to these results, demonstrating visual cortex activation during semantic tasks (Burton, Diamond, & McDermott, 2003; Noppeney, Friston, & Price, 2003), speech processing (Röder, Stock, Bien, Neville, & Rösler, 2002), and Braille reading (Sadato et al., 1996). Back-projections in the sensory processing pathways, as well as bidirectional intracortical and cortico-subcortical connections more generally, further undermine the case for informational encapsulation (for discussion, see Callaway, 2005; Churchland, Ramachandran, & Sejnowski, 1994). Given the routine violation of informational encapsulation even in simple, low-level perceptual cases, claims for the modularity of higher cognitive processes—including morality—are slimmer still.

Neurobiology and Domain Specificity

Language was the great hope for a complex cognitive ability sustained by a domain-specific, innate learning module (see, e.g., Pinker, 1994). Nonetheless, over the years, this hope has not garnered substantial support from the relevant neurobiology (Evans & Levinson, 2009; Panksepp & Panksepp, 2000), or from genetics. As for modular organization in the brain regions subserving the everyday function of language (as opposed to its acquisition)—what Haidt might call a second-order module (for language)—the case for a "grammar box" in Broca's area and a "semantic box" in Wernicke's area seems to have become weaker rather than stronger over time. Imaging studies of language comprehension and production, for example, show activity in a wide range of brain areas, including the right hemisphere and limbic structures, and careful lesion studies show results consistent with the imaging results (see Proverbio, Crotti, Zani, & Adorni, 2009; Prat, Keller, & Just, 2007; Dronkers, Wilkins, van Valin, Redfern, & Jaeger, 2004). Many questions do, of course, remain open concerning language learning and use, but at this stage, the hypothesis of a domain-specific, informationally encapsulated language module looks less promising than approaches investigating the various contributions of diverse regions.[26]

What about moral learning, in the form of MFT's learning modules and the second-order modules they produce? There is, of course, plenty of learning, of moral norms and much else; humans are prodigious learners. But what do we know about the brain that would suggest that any of this learning is subserved by domain-specific modules? Although it is well-known that there is regional specialization in the mature brain, the precise ways in which specialization and massive convergence work together is not well-understood (Meyer & Damasio, 2009; Logothetis, 2008). Furthermore, exactly how specialization emerges and the nature of the interacting roles of genes and experience-dependent plasticity are also still unclear. A leading hypothesis (Ringach, 2009; see also Dragoi, Turcu, & Sur, 2001 and Ben-Yishai, Bar-Or, & Sompolinsky, 1995) is that the statistics of the input, together with the general principle that neurons that "fire together wire together" (i.e., Hebbian learning), may account for quite a lot of observed regional specialization. This does not take us very far in the direction Haidt is going, however, as regional specialization is consistent with cross-talk and convergence at large spatial scales, and says nothing in support of modularity in the sense Haidt favors.

Likewise, small-world architecture suggests that some form of modularity may exist in densely connected micronetworks (Bullmore & Sporns, 2009; Sporns, Chialvo, Kaiser, & Hilgetag, 2004), but there is nothing to imply that modularity in *this* sense connects in any meaningful way with Haidt's domain-specific, informationally encapsulated moral modules. As noted above, the brain's small-world architecture actually weighs *against* the idea of encapsulation for complex functions owing, among other things, to the short synaptic distances between any two areas.

A further, and even more telling, point is this: Social skills, capacities, learning, and behavior can be very productively approached in terms of learning processes subserved by neurobiological systems that cut across a variety of domains (see Squire et al., 2008, Chapters 14–22). These circuits may give rise to certain patterns of representation and behavior in the social domain, but they are not modules in any sense that will give aid and comfort to Haidt; they are not *dedicated* to a single job, and they are not *encapsulated* from top–down, lateral, or subcortical input. Instead of modules, there seem to be a number of more general functions that are recruited in social learning and behavior, including pattern recognition, reward- and affect-based learning, attachment/bonding, fear, defensive behavior, impulse control, and planning (Churchland, 2011; Suhler & Churchland, forthcoming). For example, early damage to ventromedial prefrontal cortex can cause profound deficits in an individual's ability to acquire normal moral patterns of behavior (Anderson, Bechara, Damasio, Tranel, & Damasio, 1999). However, this does not mean that ventromedial prefrontal cortex is a module for moral learning (or for any other particular domain); rather, it is crucial to affect-based learning, motivation, planning, and choice across a variety of domains (see Damasio, 1994, 1996).

The availability of nonmodular neurobiological explanations for the range of abilities supposedly underwritten by second-order modules extends the challenge. Haidt characterizes these modules as "little bits of input–output programming" (Haidt & Joseph, 2007, p. 379). However, computationally speaking, having a dedicated bit of neural "programming" for each arbitrarily specific situation–

response pattern would be a highly inefficient use of neurobiological resources, and it is probably not what happens in the brain. Consider the example of much-practiced skills, a category which plausibly includes moral judgment and decision-making (Casebeer & Churchland, 2003). Here, although neurobiological changes are observed in the network involved in the skill (e.g., increases in efficiency; Wu, Kansaku, & Hallett, 2004), the components of the network do not become specifically dedicated to the task in question, instead participating in networks involved in other tasks. The pervasiveness of such neural overlap and connectivity casts doubt on the hyperspecialized modules that Haidt appears to favor with his second-order modules.

Another example of a neurobiological system that cuts across domains—including the domains picked out by the five foundations—is the mammalian system for attachment and bonding, in which the neuropeptides oxytocin and vasopressin are known to play a crucial role (for reviews, see Carter, Grippo, Pournajafi-Nazarloo, Ruscio, & Porges, 2008; Donaldson & Young, 2008). This system is deeply involved in care for preferred conspecifics such as offspring and mates (Carter et al., 2008; Carter, 2003; Carter, DeVries, & Getz, 1995), but it also plays an important role in group-oriented behaviors such as defensive responses to threats (Carter et al., 2008; Landgraf & Neumann, 2004; Carter, 1998). As such, it plausibly cuts across the harm/care and ingroup/loyalty foundations. And as indicated by recent work by experimental economists on the effects of oxytocin on exchange behavior in humans (e.g., Zak, Stanton, & Ahmadi, 2007; Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005; Zak, Kurzban, & Matzner, 2005), the oxytocin/vasopressin system may also make important contributions to behaviors associated with the fairness/reciprocity foundation.[27] Moreover, there is now evidence associating variation in a gene for the receptor for oxytocin with differences in brain structure and responsivity in the amygdala and hypothalamus and thence to temperamental variations in sociability (Tost et al., 2010). The gene should *not* be characterized as a gene "for" sociability as it is only part of a gene–brain–behavior network (Kendler & Greenspan, 2006). Nevertheless, the research is an example of what needs to be done to begin to link social behavior to neurobiology and genetics (for further discussion, see Churchland, 2011). Finally, consider that, in mammals, the neural systems regulating negative affect, including pain, appear to have been modified to support separation distress, thereby providing a platform for social emotions that are important—but certainly *not* specific—to morality (Tucker, Luu, & Derryberry, 2005; Panksepp, 1998). One emotion of this sort is the empathic experience of another individual's pain (Tucker et al., 2005). But the neural systems for pain may also be important to the psychological pain experienced when one is cheated or contemplates cheating another individual (fairness/reciprocity), and perhaps even to negative social experiences associated with violations in all of the alleged five foundational domains (see, e.g., Eisenberger, Lieberman, & Williams, 2003). Interestingly, individuals exhibit greater empathic responses when observing harms to members of their own group than when observing identical harms to outgroup individuals (Xu, Zuo, Wang, & Han, 2009), and emotions experienced upon cheating or being cheated during an economic exchange exhibit a similar ingroup intensification effect (Burton-Chellew, Ross-Gillespie, & West, 2010). How these examples may be accounted for within the evolutionary, foundational, and modularity stories told by MFT is far from apparent.

A common theme running through these examples is that evolution, as well as brain development, tends to be very conservative, typically extending, reappropriating, or modifying extant neurobiological mechanisms rather than engineering wholly new, dedicated mechanisms (e.g., modules) for each new challenge that the environment throws an organism's way (see Flint et al., 2010). These nonmodular mechanisms do, of course, help to explain the concerns that the five foundations collectively capture, but they also suggest that claims of domain specificity or of strict divisions between the foundations are not honored by the neurobiology. And although neurobiology certainly does constrain and guide the development of morality as well as the various abilities and domains of concern that comprise it, this guidance appears neither to involve modules nor to amount to innateness in any explanatorily useful sense.

CONCLUSION

Advances in genetics have encouraged research on the links between genes and behavior, while at the same time warning ever more loudly that the linkages wind their way through dynamical, intertwining networks of gene–gene, gene–brain, and gene–environment interactions. If this complexity is daunting in the case of aggression in the fruit fly—and it is (Flint et al., 2010; Dierick & Greenspan, 2006)—then it is all the more so in the case of the social behavior of large-brained mammals. To make matters worse, the remarkable plasticity of the human nervous system must also be factored in; our brains did not evolve to read books or ride bicycles, but read and ride we do. Although the general idea that our evolved biology contributes to "our nature" is unsullied, the constraints on tenable hypotheses regarding specific links between genes, brains, and behavior have become vastly more demanding in the last decade. Claims to the effect that a given behavior is "innate," "prepared for," or "organized in advance of experience" are much more difficult to substantiate, now that we have a clearer idea what, evidentially, we are up against.

Haidt's MFT seeks to explain morality by connecting evolution, psychology, and development. Although we applaud the project's ambition, the overall execution is disappointingly insensitive to the height of the evidence bar. No detailed factual support from neuroscience, molecular biology, or evolutionary biology (save for very general adaptationist speculations) is marshaled for the theory's

substantive claims, and although some speculations are consistent with what is known, others are not. Mere consistency, in any case, is a far cry from the confirmation or disconfirmation yielded in a tough test of a theory, a point long emphasized by philosophers of science (e.g., Popper, 1963).

Acknowledgments

We thank Lucia Foglia for insightful discussion of these issues. We also thank Jonathan Haidt and two anonymous referees for their helpful comments on an earlier draft.

Reprint requests should be sent to Christopher L. Suhler, Department of Philosophy, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0119, or via e-mail: csuhler@ucsd.edu.

Notes

1. For a review of scientific research on morality, with a focus on neuroscience, see Suhler and Churchland (forthcoming).
2. For short, we shall henceforth say "Haidt" rather than "Haidt and colleagues," but this should not be taken to diminish the contributions that these collaborators have made to the development of MFT. Craig Joseph, in particular, has been the coauthor of all of Haidt's works focusing on innateness and modularity, two concepts that will be central to our discussion in what follows.
3. Haidt and Joseph (2004) actually acknowledge only four "official" foundations while mentioning a strong candidate for a fifth foundation in a footnote (p. 63, note 15). That candidate—ingroup/loyalty—has, in subsequent works (e.g., Haidt & Graham, 2009; Haidt & Joseph, 2007), been elevated to the status of a full-fledged foundation.
4. This list is adapted from Haidt and Joseph (2007, p. 382, Table 19.1).
5. The quote Haidt and Joseph attribute to Sperber is, in fact, slightly inaccurate. Sperber, himself borrowing a phrase from Fodor (1983, p. 37), actually says "modular to some interesting *extent*" (1994, pp. 39, 48, emphasis added). We note this for the sake of accuracy, but of course, it does not affect the substance of the idea being appealed to.
6. This role for learning is a major way in which MFT incorporates constructivist as well as nativist elements. Our focus here will be predominantly on the nativist elements, for a number of reasons. First, the nativist portion of the theory is the one which has been most extensively developed, for instance, though arguments concerning selection for certain domains of concern and the positing of specific mechanisms such as modules. Second, it is the part of the theory that we believe to be most questionable in its details. Third, and relatedly, without an adequate understanding of what sorts of evolutionary and cognitive processes are constraining moral development, efforts to conduct developmental studies that flesh out the theory's constructivist elements are likely to founder upon a lack of specificity in the theoretical edifice underlying them. For example, without a clear understanding of what the purported "learning modules" are and what computational features they possess, the design and interpretation of developmental studies will be seriously underdetermined.
7. Notably, the modules that develop in this way need not always be concerned with social patterns, events, and so forth which directly correspond to the evolutionary challenges listed earlier. Consequently, one output of the learning module for fairness/reciprocity might be a "cutting-in-line detector" (Haidt & Joseph, 2007, p. 379), even though forming queues was not, in anything like its modern form, a part of the social environment in which the bulk of hominid evolution took place. In allowing for this possibility, Haidt and Joseph (2007, p. 381) invoke the distinction, common in the modularity literature (e.g., Sperber, 1994), between a module's *proper domain* and its *actual domain*. The proper domain is the set of triggers that natural selection shaped the module to respond to, whereas the actual domain is the set of all events that actually do trigger the module. For fairness/reciprocity, then, an example of a trigger falling within the module's proper (and actual) domain would be someone cheating in an exchange of goods or favors, whereas an example of a trigger falling in its actual (but *not* proper) domain would be someone cutting in front of you in line at the store. This distinction allows MFT to account for the fact that our current moral concerns extend well beyond situations that have direct counterparts in our evolutionary past.
8. We do note that in psychology and neuroscience—but not, generally speaking, in philosophy (although there have long been important dissenters—e.g., Hume, 1739/1740/2008)—researchers are moving away from this reasoning-focused picture as evidence accumulates that affect is central to moral judgment (for reviews, see Suhler & Churchland, forthcoming; Haidt, 2007; Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005).
9. For others who diverge from the "unitary principle" approach, see, e.g., Flanagan and Williams (2010), Hursthouse (1999), Johnson (1993), Flanagan (1991), and MacIntyre (1981).
10. Although it has fallen out of favor among many contemporary moral philosophers, as well as many scientists studying morality, this feature of the theory, too, has a long history. It dates at least to ancient Greece, where Aristotle, in his *Nicomachean Ethics* (see the 2000 translation by Roger Crisp), devotes much attention to describing not only the virtues one must possess but also the sort of social and political environment in which the youth must be raised if they are to develop those virtues. For more recent examples of work that incorporates evolutionary as well as developmental elements, see Baumeister (2005), Hrdy (1999), and Ridley (1997).
11. Haidt professes to be sensitive to this, noting, in the course of describing the notion of innateness at work in his theory, the importance of being clear about how one is using this term (see Haidt & Joseph, 2007, pp. 373–374).
12. See also Richardson (2007) for a critical discussion of the concept of preparedness and the supposedly paradigmatic examples (e.g., fear of snakes and spiders) around which the concept has been built.
13. See, e.g., Buller and Hardcastle (2007), as well as the Neurobiology and MFT section of this article.
14. See Richerson and Boyd (2005), as well as Richardson (2007), for more on the problems with the move from a trait's universality to claims that it must be genetically hardwired.
15. Perhaps the most famous example of this in social psychology is the vast body of research on cognitive dissonance initiated by Festinger's (1957) seminal work. It is worth nothing that mainstream psychological researchers studying cognitive dissonance and numerous other phenomena exhibiting some degree of (what Haidt would regard as) partial informational encapsulation have rarely, if ever, been tempted to make the move from this partial encapsulation to conclusions about modularity.
16. Note also that adding the requirement that some supposedly modular trait has been shaped by natural selection would do little to save the account of modularity on offer. Although certain evolutionary psychologists (e.g., Cosmides & Tooby, 1994, 1997) have tried to argue a priori that natural selection "must" lead to mental modules, there is little or no actual evidence that it really does lead to modular cognitive organization (see the Neurobiology and MFT section). As such, one cannot move from claims about adaptiveness to claims about modularity. Indeed, it is entirely possible to construct a scientific account of morality that sees an important role for evolution but makes no mention of modular processes (see Churchland, 2011; Suhler & Churchland, forthcoming).

17. Haidt is careful to leave open this possibility, at times saying that morality consists of "at least" five foundations (see, e.g., Haidt & Graham, 2009, p. 381).

18. We are not, to be clear, arguing that industry and modesty *are* foundational values, only that they have as good a claim to be foundations as those in Haidt's current list. Other concerns may have strong claims as well; *patience* and *humor*, for example, have figured in other lists.

19. Haidt and Joseph may tacitly acknowledge this when they list as a characteristic emotion of the ingroup/loyalty foundation "rage at traitors" (2007, p. 382, Table 19.1); the use of "rage" seems like little more than linguistic sleight of hand to avoiding mentioning anger as a characteristic emotion for two different foundations.

20. Similar points about content can be made about other foundations. For example, different groups may have radically different beliefs about what is "harmful/fair." American libertarians, for instance, may equate fairness with whatever results from social and economic activity that is free, to the greatest extent possible, from regulation or other forms of government involvement, and regard departures from this as intolerably unfair and harmful. By contrast, American liberals, as well as many western Europeans, may see the libertarian's ideal of fairness as grossly *unfair*, instead endorsing a conception of fairness in which progressive redistribution of wealth and state involvement in certain services is a requirement of any fair social arrangement.

21. As in our discussion of additional foundations in the Additional Foundations section, we adopt the framework and language of MFT in this section merely as a device for highlighting phenomena we believe MFT's current foundational taxonomy does not adequately account for. As such, talk of (say) how areas of concern might be the product of one or another foundation should not be taken as an endorsement of MFT or of the particular foundations it posits.

22. In a recent paper, Haidt, Graham, and Joseph (2009) argue for this very conclusion. Using cluster analysis, Haidt and his coauthors find four principal moral subtypes characterized by different combinations of foundational concerns. In addition to the familiar "secular liberal" and "social conservative" subtypes, the analysis yields two other subtypes, which the researchers term "libertarian" and "the religious left." As we describe in this section and the next, libertarians are one group we have concerns about MFT's present ability to account for, and so we look forward to seeing the results of Haidt et al.'s (2009) analysis further incorporated into the theory in their future work.

23. Even this broader survey of "modern" groups would leave out a good deal of complexity, as it neglects historical groups that do not fit neatly into the contemporary categories of "modern" and "traditional". How, for example, would the moralities of the ancient Greeks or of members of the European Enlightenment be situated within a foundations-style framework?

24. Haidt's learning modules, in particular, cry out for a more detailed neurobiological explanation, because from what Haidt has said about them, they appear to be far less the products of experience than second-order modules. They seem, in other words, to be the clearest example in MFT of something constituting "organization in advance of experience."

25. For data on the effects of prefrontal activity on sensory processing, see Duncan (2001), as well as Miller and Cohen (2001). See also Callaway (2005) on the substrates for interaction between parallel pathways *within* V1.

26. Incidentally, a dedicated face recognition area in the fusiform gyrus is perhaps a more probable candidate than language for modularity. Although Nancy Kanwisher (2010) has marshaled evidence from neuroimaging and lesions in support of this hypothesis, even here, the case remains controversial (see Hanson & Halchenko, 2008, as well as the collection by Hanson & Bunzl, 2010).

27. For a more detailed account of the possible relationship between the oxytocin/vasopressin system and morality, see Suhler and Churchland (forthcoming), as well as Churchland (2011).

REFERENCES

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience, 2,* 1032–1037.

Aristotle. (2000). *Nicomachean ethics* (R. Crisp, Trans.). Cambridge, UK: Cambridge University Press.

Baumeister, R. F. (2005). *The cultural animal: Human nature, meaning, and social life*. New York: Oxford University Press.

Ben-Yishai, R., Bar-Or, R. L., & Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences, U.S.A., 92,* 3844–3848.

Blackburn, S. (2008). *How to read Hume*. London: Granta.

Bowles, S. (2008). Conflict: Altruism's midwife. *Nature, 456,* 326–327.

Brooks, D. (2009, April 6). The end of philosophy. *The New York Times*. www.nytimes.com. Accession date: 28 July 2010.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences, 1124,* 1–38.

Buller, D. J., & Hardcastle, V. G. (2007). Evolutionary psychology, meet developmental neurobiology: Against promiscuous modularity. *Brain and Mind, 1,* 307–325.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience, 10,* 186–198.

Burton, H., Diamond, J. B., & McDermott, K. B. (2003). Dissociating cortical regions activated by semantic and phonological tasks: A fMRI study in blind and sighted people. *Journal of Neurophysiology, 90,* 1965–1982.

Burton-Chellew, M. N., Ross-Gillespie, A., & West, S. A. (2010). Cooperation in humans: Competition between groups and proximate emotions. *Evolution and Human Behavior, 31,* 104–108.

Buzsáki, G. (2006). *Rhythms of the brain*. New York: Oxford University Press.

Callaway, E. M. (2005). Structure and function of parallel pathways in the primate early visual system. *Journal of Physiology, 566,* 13–19.

Carter, C. S. (1998). Neuroendocrine perspectives on social attachment and love. *Psychoneuroendocrinology, 23,* 779–818.

Carter, C. S. (2003). Developmental consequences of oxytocin. *Physiology & Behavior, 79,* 383–397.

Carter, C. S., DeVries, A. C., & Getz, L. L. (1995). Physiological substrates of mammalian monogamy: The Prairie Vole model. *Neuroscience and Biobehavioral Reviews, 19,* 303–314.

Carter, C. S., Grippo, A. J., Pournajafi-Nazarloo, H., Ruscio, M. G. & Porges, S. W. (2008). Oxytocin, vasopressin and sociality. In I. D. Neumann & R. Landgraf (Eds.), *Progress in brain research 170: Advances in vasopressin and oxytocin: From genes to behaviour to disease* (pp. 331–336). New York: Elsevier.

Casagrande, V. A., Sáry, G., Royal, D., & Ruiz, O. (2005). On the impact of attention and motor planning on the lateral geniculate nucleus. *Progress in Brain Research, 149,* 11–29.

Casebeer, W. D. (2003). *Natural ethical facts: Evolution, connectionism, and moral cognition*. Cambridge, MA: MIT Press.

Casebeer, W. D., & Churchland, P. S. (2003). The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy, 18,* 169–194.
Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton: Princeton University Press.
Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch & J. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23–60). Cambridge, MA: MIT Press.
Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 85–116). Cambridge: Cambridge University Press.
Cosmides, L., & Tooby, J. (1997). The modular nature of human intelligence. In A. Scheibel & J. W. Schopf (Eds.), *The origin and evolution of intelligence* (pp. 71–101). Boston: Jones and Bartlett.
Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: G.P. Putnam and Sons.
Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, 351,* 1413–1420.
de Quervain, J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305,* 1254–1258.
Dierick, H. A., & Greenspan, R. J. (2006). Molecular analysis of flies selected for aggressive behavior. *Nature Genetics, 38,* 1023–1031.
Donaldson, Z. R., & Young, L. J. (2008). Oxytocin, vasopressin, and the neurogenetics of sociality. *Science, 322,* 900–904.
Dragoi, V., Turcu, C. M., & Sur, M. (2001). Stability of cortical responses and the statistics of natural scenes. *Neuron, 32,* 1181–1192.
Dronkers, N. F., Wilkins, D. P., van Valin, R. D., Jr., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition, 92,* 145–177.
Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience, 2,* 820–829.
Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302,* 290–292.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32,* 429–448.
Farris, G. (2009, February 9). PETA dresses in KKK garb outside Westminster Dog Show. *USA Today*. Retrieved from www.usatoday.com. Accession date: 28 July 2010.
Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row.
Flanagan, O., & Williams, R. A. (2010). What does the modularity of morals have to do with ethics? Four moral sprouts plus or minus a few. *Topics in Cognitive Science, 2,* 430–453.
Flanagan, O. J. (1991). *Varieties of moral personality: Ethics and psychological realism*. Cambridge, MA: Harvard University Press.
Flint, J., Greenspan, R. J., & Kendler, K. S. (2010). *How genes influence behavior*. New York: Oxford University Press.
Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.
Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24,* 153–172.
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029–1046.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44,* 389–400.
Greene, J. D., Sommerville, R. B., Nystrom, L., Darley, J., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108.
Greenspan, R. J. (2001). The flexible genome. *Nature Reviews Genetics, 2,* 383–387.
Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience, 2,* 685–694.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–834.
Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316,* 998–1002.
Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In J. Jost, A. C. Kay, & H. Thorisdottir (Eds.), *Social and psychological bases of ideology and system justification* (pp. 371–401). New York: Oxford University Press.
Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory, 20,* 110–119.
Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133,* 55–66.
Haidt, J., & Joseph, C. (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind, Volume 3: Foundations and the future* (pp. 367–392). New York: Oxford University Press.
Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613–628.
Hanson, S. J., & Bunzl, M. (2010). *Foundational issues in human brain mapping*. Cambridge, MA: MIT Press.
Hanson, S. J., & Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Computation, 20,* 486–503.
Hrdy, S. B. (1999). *Mother nature: A history of mothers, infants, and natural selection*. New York: Pantheon Books.
Hume, D. (1739/1740/2008). In D. F. Norton & M. J. Norton (Eds.), *A treatise of human nature*. Oxford: Oxford University Press.
Hursthouse, R. (1999). *On virtue ethics*. Oxford: Oxford University Press.
Johnson, M. (1993). *Moral imagination: Implications of cognitive science for ethics*. Chicago: University of Chicago Press.
Kant, I. (1785/1998). *Groundwork of the metaphysics of morals*. Cambridge: Cambridge University Press.
Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 11163–11170.

Kendler, K. S., & Greenspan, R. J. (2006). The nature of genetic influences on behavior: Lessons from "simpler" organisms. *American Journal of Psychiatry, 163,* 1683–1694.
Kohlberg, L. (1981). *The philosophy of moral development: Moral stages and the idea of justice*. San Francisco, CA: Harper & Row.
Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages*. San Francisco: Harper & Row.
Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435,* 673–676.
Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience, 2,* 635–642.
Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin, 129,* 723–746.
Landgraf, R., & Neumann, I. D. (2004). Vasopressin and oxytocin release within the brain: A dynamic concept of multiple and variable modes of neuropeptide communication. *Frontiers in Neuroendocrinology, 25,* 150–176.
Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453,* 869–878.
MacIntyre, A. C. (1981). *After virtue: A study in moral theory*. Notre Dame, IN: University of Notre Dame Press.
Maimon, G., Straw, A. D., & Dickinson, M. H. (2010). Active flight increases the gain of visual motion processing in Drosophila. *Nature Neuroscience, 13,* 393–399.
Marcus, G. F. (2004). *The birth of the mind: How a tiny number of genes creates the complexities of human thought*. New York: Basic Books.
Mellet, E., Tzourio-Mazoyer, N., Bricogne, S., Mazoyer, B., Kosslyn, S. M., & Denis, M. (2000). Functional anatomy of high-resolution visual mental imagery. *Journal of Cognitive Neuroscience, 12,* 98–109.
Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neurosciences, 32,* 376–382.
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202.
Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6,* 799–809.
Niell, C. M., & Stryker, M. P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron, 65,* 472–479.
Noppeney, U., Friston, K. J., & Price, C. J. (2003). Effects of visual deprivation on the organization of the semantic system. *Brain, 126,* 1620–1627.
Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford University Press.
Panksepp, J., & Panksepp, J. B. (2000). The seven sins of evolutionary psychology. *Evolution and Cognition, 6,* 108–131.
Paradiso, M. A., MacEvoy, S. P., Huang, X., & Blau, S. (2005). The importance of modulatory input for V1 activity and perception. *Progress in Brain Research, 149,* 257–267.
Pinker, S. (1994). *The language instinct* (1st ed.). New York: W. Morrow and Co.
Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.
Prat, C. S., Keller, T. A., & Just, M. A. (2007). Individual differences in sentence comprehension: A functional magnetic resonance imaging investigation of syntactic and lexical processing demands. *Journal of Cognitive Neuroscience, 19,* 1950–1963.
Proverbio, A. M., Crotti, N., Zani, A., & Adorni, R. (2009). The role of left and right hemispheres in the comprehension of idiomatic language: An electrical neuroimaging study. *BMC Neuroscience, 10,* 116.
Richardson, R. C. (2007). *Evolutionary psychology as maladapted psychology*. Cambridge, MA: MIT Press.
Richerson, P. J., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.
Ridley, M. (1997). *The origins of virtue: Human instincts and the evolution of cooperation*. New York: Viking.
Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, K. D. (2002). A neural basis for social cooperation. *Neuron, 35,* 395–405.
Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *NeuroReport, 15,* 2539–2543.
Ringach, D. L. (2009). Spontaneous and driven cortical activity: Implications for computation. *Current Opinion in Neurobiology, 19,* 439–444.
Röder, B., Stock, O., Bien, S., Neville, H., & Rösler, F. (2002). Speech processing activates visual cortex in congenitally blind humans. *European Journal of Neuroscience, 16,* 930–936.
Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32,* 880–892.
Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M. P., Dold, G., et al. (1996). Activation of the primary visual cortex by Braille reading in blind subjects. *Nature, 380,* 526–528.
Sherman, S. M. (2005). Thalamic relays and cortical functioning. *Progress in Brain Research, 149,* 107–126.
Shermer, M. (2009). Political science. *Scientific American, 301,* 38.
Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, and divinity), and the "big three" explanations of suffering. In A. Brandt & P. Rozin (Eds.), *Morality and health* (pp. 119–169). New York: Routledge.
Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). Cambridge: Cambridge University Press.
Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (pp. 53–68). New York: Oxford University Press.
Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences, 8,* 418–425.
Squire, L. R., Berg, D., Bloom, F. E., du Lac, S., Ghosh, A., & Spitzer, N. C. (2008). *Fundamental neuroscience* (3rd ed.). Boston: Academic Press.
Suhler, C. L., & Churchland, P. S. (forthcoming). The neurobiological basis of morality. In J. Illes & B. J. Sahakian (Eds.), *The Oxford handbook of neuroethics*. Oxford: Oxford University Press.
Super, H., van der Togt, C., Spekreijse, H., & Lamme, V. A. F. (2003). Internal state of monkey primary visual cortex (V1) predicts figure–ground perception. *Journal of Neuroscience, 23,* 3407–3414.

Tost, H., Kolachana, B., Hakimi, S., Lemaitre, H., Verchinski, B. A., Mattay, V. S., et al. (2010). A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic–limbic structure and function. *Proceedings of the National Academy of Sciences, U.S.A., 107,* 13936–13941.

Tucker, D. M., Luu, P., & Derryberry, D. (2005). Love hurts: The evolution of empathic concern through the encephalization of nociceptive capacity. *Developmental Psychopathology, 17,* 699–713.

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge: Cambridge University Press.

Wade, N. (2007, September 18). Is "do unto others" written into our genes?. *The New York Times.* Retrieved from www.nytimes.com. Accession date: 28 July 2010.

Wu, T., Kansaku, K., & Hallett, M. (2004). How self-initiated memorized movements become automatic: A functional MRI study. *Journal of Neurophysiology, 91,* 1690–1698.

Xu, X., Zuo, X., Wang, X., & Han, S. (2009). Do you feel my pain? Racial group membership modulates empathic neural responses. *Journal of Neuroscience, 29,* 8525–8529.

Zak, P. J., Kurzban, R., & Matzner, W. T. (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behavior, 48,* 522–527.

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS ONE, 2,* e1128.

How Moral Foundations Theory Succeeded in Building on Sand: A Response to Suhler and Churchland

Jonathan Haidt and Craig Joseph

Suppose you are an architect and you have recently completed a challenging project: designing and building a sturdy modern house on a sandy stretch of ground where several previous architects had failed. The shifting ground had cracked their one-piece rigid concrete foundations. You vowed not to repeat their mistakes, so you designed a novel foundational system that avoided the use of concrete altogether. You drove steel rods down into rockier soil, created five independent platforms to support five modular units, and then linked the units together with short flexible corridors. You left plenty of room for expansion—the modular design makes it easy for the homeowner to add additional units as needed.

The initial reviews of your modular house are excellent, and other architects begin applying your technique, with good results.[1] Imagine your trepidation, then, when a major architectural critic writes a review entitled "A foundation built on sand?", in which she warns that your house will soon collapse and that your project is useful primarily as an object lesson in what *not* to do.

You begin reading the review. It starts off with an extremely accurate summary of the design challenges you faced and of the innovative ways that you met those challenges. It praises you for having solved four of the major problems that doomed previous attempts to build on this sandy ground. (You are grateful for this praise.) So imagine your confusion as you continue to read and discover that your critic's three major complaints are as follows:

(1) Your steel rods are not strong enough to support the house (when in fact the house is already standing).
(2) There is no garage (which is true, but a strength of your design is that future owners can easily add whatever rooms or structures are needed).
(3) You failed to extend your steel rods down to the center of the earth (which is true, but it is both impossible and unnecessary to do so).

These three complaints are close analogues of the complaints Suhler and Churchland (2011) level against Moral Foundations Theory (MFT): (1) Our concepts of innateness and modularity are defective and cannot support the theory. (2) There are additional candidates for foundationhood. (3) We failed to link MFT to neuroscience and genetics. In this essay, we will speak for the team that designed MFT and oversees its ongoing testing and revision; the team includes Peter Ditto, Jesse Graham, Ravi Iyer, Sena Koleva, and Brian Nosek. We will first address Complaints 2 and 3, which are true statements that are not valid criticisms. We will then address Complaint 1, which is a more substantive charge.

University of Virginia

COMPLAINT 2: THERE IS NO GARAGE

> Second, both the theory's proposed number of moral foundations and its taxonomy of the moral domain appear contrived, ignoring equally good candidate foundations and the possibility of substantial intergroup differences in the foundations' contents (Suhler & Churchland, 2011, p. 2103).

We have said from the beginning (Haidt & Joseph, 2004) that our list of proposed foundations was a starting point, not an exhaustive list. MFT was an attempt to specify the best candidates, the best spots at which to bridge the topics discussed by evolutionary psychologists (e.g., reciprocal altruism and coalitional psychology) with phenomena described by anthropologists (e.g., reciprocal gift-giving and tribalism). We proposed our list (see Haidt & Graham, 2007) and then posted a challenge at www.MoralFoundations.org. We offered to pay $1000 to anyone who could show that additional foundations were needed or that the current foundations should be rearranged. We received 15 challenges and have collected data to test several of them so far. We are now in the process of revising the theory and are likely to add a foundation related to liberty or domination, for which the evolutionary story has been told by Chris Boehm (1999). It includes the hypervigilance of egalitarian hunter–gatherers for any sign of alpha male behavior, including boasting. (This new foundation will, therefore, support Suhler and Churchland's intuition that there is something widely disliked about boasting.) We are also investigating a foundation related to wastefulness, and we are considering revising the fairness foundation to exclude equality and focus on equity, which would support intuitions related to the Protestant Work Ethic and concerns about industry (e.g., slackers

and freeloaders who want to be a part of the group but do not contribute their fair share—one of the principal concerns of today's Tea Partiers).

In other words, MFT was designed to be revisable, and it is being revised. It simply cannot be a complaint against MFT that we did not start with the final list of foundations. If all scientists took Suhler and Churchland's approach to theory construction, there would be few new theories.

There is a larger scientific issue at stake here. Suhler and Churchland accuse us of an "ad hoc" approach to theory construction, and they advise us to take a more "principled" approach. But the "principled" approach is part of what doomed previous grand theories in psychology (e.g., Kohlberg, 1969). If you start by fixating on a principle (e.g., that morality is justice, or empathy, or harm reduction, or prosocial behavior) and then develop your theory in a logical way on the basis of that principle, you will construct an elegant and parsimonious theory, but it will crack under the weight of empirical data (like the one-piece concrete foundations in our opening metaphor).

Elsewhere, one of us (Haidt, in press) is developing Hume's claim that morality is like taste, not like reasoning. Imagine if taste scientists had been told that it was "ad hoc" to create a theory of taste by looking at the tongue and trying to figure out how many different taste receptors it has. Shouldn't taste scientists proceed in a more principled way, such as by analyzing the nutritional needs of human beings and then positing a set of receptors that would guide people to the right foods? And doesn't the recent discovery of a fifth taste receptor (umami or glutamate) show that the initially ad hoc list of four taste receptors was a failure? No. There is no a priori or principled way to figure out how taste works. You need to look at the tongue, pick the best candidates, and let your fellow scientists show you what you missed. That is what we did for moral psychology. We reject on principle the idea that moral psychology should proceed in a principled (rather than descriptive, naturalistic) way, and that it should value parsimony above explanatory adequacy.

As for the claim that liberals sometimes rely upon the purity foundation, particularly with regard to environmental purity: we agree. We never said that any group lacks access to any of the foundations. Our claims have always been about relative reliance upon each foundation, and we have found, using many kinds of questions and question formats, that social conservatives (on average) live in a world more saturated with the magical thinking of the purity foundation than do liberals. Liberals score lower on measures of disgust sensitivity that have nothing to do with politics (Inbar, Pizarro, & Bloom, 2009). Just compare the writings of Peter Singer (1979), who says that nothing is sacred and all must be evaluated consequentially, to Leon Kass (1997), who says "shallow are the souls who have forgotten how to shudder." We think that there is a real difference here and that MFT captures that difference neatly. Suhler and Churchland posit that if we were to measure a broader range of content, "the gap between conservatives' and liberals' concern with *this* foundation might well entirely disappear." We bet it would not, and as we develop more ways of measuring foundational concerns, we seem to be winning the bet (see, e.g., Graham, Haidt, & Nosek, 2009, Studies 3 and 4, which used novel methods not subject to Suhler & Churchland's concerns).

As for the claim that MFT cannot handle libertarians or others who do not fit on the left–right axis, this is easily disproved. MFT offers five dimensions with which ideologies can be characterized, allowing for far more precision than the one-dimensional left–right axis. Each foundation predicts unique variance in political attitudes, over and above people's self-placement on the left–right dimension (Koleva, Graham, Haidt, Iyer, & Ditto, submitted). Haidt, Graham, and Joseph (2009) performed a cluster analysis of participants' scores on the five foundations and discovered that libertarians are relatively low on all five dimensions, whereas communitarians are relatively high on all five. These two profiles contrasted with those of liberals (high on the first two and low on the last three) and conservatives (low, relative to other groups, on the first two and high on the last three). More recently, Iyer, Koleva, Graham, Ditto, and Haidt (submitted) compiled the most extensive psychological profile of libertarians ever assembled, showing dozens of ways in which libertarians differ from liberals and from conservatives (who often resemble each other much more than they resemble libertarians). The key difference is that libertarians hold almost nothing sacred, with the exception of liberty (our new liberty or domination foundation).

In summary, Suhler and Churchland are correct that we did not build a garage on the initial house, but our modular design allows us add one. We have added one and are getting a lot of use out of it. We are looking forward to future expansions too.

COMPLAINT 3: YOU FAILED TO EXTEND YOUR STEEL RODS DOWN TO THE CENTER OF THE EARTH

> Third, the mechanisms (viz., modules) and categorical distinctions (viz., between foundations) proposed by the theory are not consilient with discoveries in contemporary neuroscience concerning the organization, functioning, and development of the brain (Suhler & Churchland, 2011, p. 2103).

Suhler and Churchland assert that "innateness hypotheses are now expected to be supported by, or at least consilient with evidence from" developmental psychology, neurobiology, and genetics. We are surprised to hear that this is now a common expectation. Of course, innateness hypotheses should not be *incompatible* with well-established findings from those fields, but Suhler and Churchland are asking for much more; they want to see positive links to those three fields, including the identification of candidate genes and neural systems. Such

positive linkage with developmental psychology is reasonable enough; as cultural psychologists, we set as one of our main design challenges the need to create a theory that would explain the divergent developmental paths taken by children in diverse cultures. (See Haidt & Joseph, 2004, 2007, on the development of virtues; Suhler and Churchland praise us on this point.)

But *genetics*? One of the biggest news stories in science in the last few years has been that, despite the fact that just about everything is heritable (Turkheimer, 2000), there do not appear to be genes "for" traits. The human genome project failed to find genes or even sets of dozens of genes that account for more than a few percent of the variance in any target disease or trait. Even for physical height, which has a heritability of 0.9 and can be measured with nearly perfect accuracy, nobody can find a gene or a set of genes that explain why some people are taller than others (Turkheimer, in press). The most successful of these genome-wide association scans identified *27 genes that, when combined, explained just 3.7% of the variance* in height (Gudbjartsson et al., 2008). What hope, then, is there for finding genes "for" reciprocity, loyalty, or authority? Of course, the genome codes for traits somehow or other, but nobody knows how. Yet, despite the disappointing news emerging from the human genome project, Suhler and Churchland claim that any scientist who proposes a nativist theory is now "expected" to identify genes that are at least associated with the innate content. This is equivalent to demanding that all new buildings must dig their foundations down to the center of the earth. It cannot be done today, it might be impossible in principle, and if it is required of all new nativist theories, then there will be no new nativist theories.

The same problem applies to neuroscience, although not as starkly. We have always treated moral modules as *functional* modules, not as physical, anatomical, or neurobiological modules. We were attracted to modularity, with partial (not complete) encapsulation, because of our observations of moral dumbfounding. For example, when asked about an adult brother and sister who have sex once, using two forms of birth control, many participants condemn the action. When pressed to justify their condemnation, many subjects search for reasons, fail to find any, and then admit that they cannot justify their condemnation. Yet, they continue to maintain that the action was wrong and are sometimes puzzled by their own continued condemnation. These situations are analogous to optical illusions, such as the Muller–Lyer illusion: One line continues to look longer, even after you measure the two lines yourself. In both cases, the judgment is partially encapsulated; it is not fully revised by the acquisition of other relevant information.

Suhler and Churchland's long section on neurobiology assumes that we are positing *neurobiological* modules—specific neural circuits that correspond to moral foundations or, at least, to the component operations that comprise moral judgment. But we are not, and we do not see how the phenomenon of moral dumbfounding (or any psychological phenomenon) can be negated (or declared "not consilient") with *any* finding about neurons and circuits. It is just too low a level of analysis, at least until we have a neuroscience so complete that we can say how neural activity fully instantiates and constrains specific moral judgments. It is interesting that neuroanatomical circuits are often loopy. But does that mean that no knowledge can be partially encapsulated? Should we inform our dumbfounded participants that they cannot be dumbfounded because their neural circuits are too loopy to allow it? Likewise, it is interesting that neurons exhibit spontaneous activity. But how can that fact make MFT (or *any* theory of higher cognition) more or less plausible?

In summary, Suhler and Churchland's third complaint is that we have made no effort to seek consilience with neuroscience and genetics. We agree with their claim but cannot see how this counts as a mark against MFT. If their "expectation" about the requirements for nativist theories were to become widespread, there would be no more nativist theories. And that, we suspect, is why they have proposed an impossibly high bar for nativist theories.

COMPLAINT 1: YOUR STEEL RODS ARE NOT STRONG ENOUGH TO SUPPORT THE HOUSE

Since the 1980s, there has been a slight correlation between geography and attitudes about nativism in the United States. The "East Pole" of this intellectual dimension has been located in the Northeast, particularly at Harvard and the Massachusetts Institute of Technology, where ideas about modularity, computational theory, and evolutionary psychology mixed together to support a nativist perspective on mind and behavior (see Pinker, 2002). The "West Pole" is in California, particularly at the University of California–Berkeley and the University of California–San Diego, where an interest in connectionism and brain plasticity has led to a preference for more empiricist (experience-based) explanations (see Elman et al., 1996). Churchland is a West Poler. That is her choice; good arguments can be marshaled on both sides. But if a pair of West Polers set an impossibly high bar for nativist theories—all nativist theories—and then declare that MFT does not meet that bar, it cannot count as a criticism of MFT specifically. It is simply a declaration of what West Polers believe.

For example, Suhler and Churchland declare that,

> to avoid mere hand-waving, innateness claims have to provide evidence that the traits they target tend to the "insensitive-to-environmental-influences" end of the spectrum, *and*, for adaptationist accounts, that these traits were selected for in the course of human evolution (p. 2105).

Because few or no psychological traits are "hard wired" or "insensitive to environmental influences" and because

it is very difficult to *prove* that a trait was selected for, Suhler and Churchland are essentially saying "bring us a colorless green idea *and* the broomstick of the Wicked Witch of the West, and only then will we certify that your theory is more than hand waving."

We have been very clear that by "innate" we mean "organized in advance of experience" (Marcus, 2004). We have consistently borrowed Marcus's metaphor that the mind is like a book. The genes write the first draft into neural tissue (although there may be no genes "for" any specific modules or for any specific paragraphs in the book). Experience (nurture) then revises the draft. Some chapters of the book are heavily edited by experience in some cultures but only lightly edited in others. Innate traits need not be visible in all known cultures. For example, the preference of most teenage boys for heterosexual rather than homosexual sex is still innate, even if some New Guinea societies are able to engineer a period of homosexuality (as described by Herdt, 1981). As long as there is some *organization in advance of the editing*, we join Marcus in calling it innate.

In a previous work (e.g., Haidt & Joseph, 2007), we have drawn on Sperber's (1994, 2005) notion of "massive" or "teeming modularity" as a way of formulating the innate part of moral functioning. Thus, we are not bothered by Suhler and Churchland's charge that our "weak" nativism may apply to "too many" cognitive and behavioral traits. Too many? Given that just about every trait you can imagine, from divorce proneness to musical preferences, is heritable, we are quite content to say that most behavioral and cognitive traits (including the moral foundations and much else) draw to some degree on innate traits, abilities, and interests. Whether we are too "promiscuous" with our nativism or they are too "prudish" depends mostly on which pole you prefer.

Suhler and Churchland also charge that our use of modularity is "murky," a "black box" amounting to little more than a "restatement of the behavioral data, lacking computational, neurobiological, or other details." We readily grant that we are not computational neuroscientists. We have not yet specified in detail exactly what is inside each module (although Haidt, in press, will give far more detail). MFT is not yet a complete theory spanning all levels of analysis, and we hope that, in time, it will be. But is the incompleteness of a theory a reason to reject it or to develop it?

Suhler and Churchland seem to have taken Fodor's (1983) theory of modularity as the gold standard for what a module is. We agree with them and with Fodor that this standard is so high that there are probably no Fodorian modules in higher cognition. According to Barrett and Kurzban (2006, p. 628), "opponents of modern views of modularity have critiqued modern positions as though the original (Fodorian) conception of modularity were intended." So let us forget Fodor modules and look at what evolutionary psychologists actually mean when they talk about modularity. The answer is simple: *functional specialization*. As Barrett and Kurzban point out, functional specialization is a basic feature of systems designed by natural selection. The digestive system, for example, is a functionally specialized module within the body, and its function is to extract nutrients from food. It, in turn, is composed of smaller modules, each with a specialized function related to the specific type of input that it receives. You cannot understand any structure in the digestive system without first knowing its function and its inputs.

The situation is similar in cognition: Different kinds of information are handled by different systems. "Functionally specialized mechanisms with formally definable informational inputs are characteristic of human (and nonhuman) cognition and ... these features should be identified as the signal properties of 'modularity'" (Barrett & Kurzban, 2006, p. 630). Applying this definition to MFT leads to this claim: The moral mind includes at least five sets of modules that are functionally specialized to handle informational inputs related to social events involving (1) care versus harm, (2) fairness versus cheating, (3) loyalty versus betrayal, (4) authority versus subversion, and (5) sanctity versus degradation. This claim may need some adjustments over time in the number and exact functions of these modules, but it is hardly a *vacuous* claim. It offers a sharp contrast with Suhler and Churchland and all other antinativist theories that try to explain moral functioning as a product of domain-general cognitive or developmental mechanisms, such as social learning. Functional modules might or might not (someday) turn out to be coincident with neurological models, but they should be evaluated and tested by research on how people process information. Which theory fits the data better, a modular theory or a general learning theory? Neither side has the right to claim to be the "conservative" answer and then to require its opponent to prove ($p < .05$) its superiority. It is a straightforward competition: Which approach better fits the facts of moral psychology?

Suhler and Churchland are correct that "the mere commonness of moral norms corresponding to the five foundations" does not indicate the existence of modules. But how would they explain otherwise weird cross-cultural similarity in the operation of rules of purity and pollution (see Haidt, 2006, Chap. 9)? How would they explain the emergence in multiple cultures, around the age of seven, of the game known in the United States as "cooties" (Samuelson, 1980)? In this game, children who are either of the opposite sex or who are low in popularity suddenly become contagious—their mere touch transfers "cooties," which, in the American version, must be treated with a (pretend) vaccine. When you find highly structured practices that are widespread across cultures and that seem to emerge even in the absence of encouragement from adults (as is the case with cooties), it becomes increasingly plausible that the behaviors did not emerge from generalized social learning. Rather, they reflect the existence of specialized modules, which make it easy to learn

norms, behaviors, and games related to contagion and purity.

We close with this example from Immanual Kant, a "systemizer" (Baron-Cohen, 2009) who built up his theory of morality in the most a priori and principled possible way. Yet, even Kant (1797/1996) found within himself an inexplicable moral horror at masturbation:

> That such an unnatural use (and so misuse) of one's sexual attributes is a violation of one's duty to himself and is certainly in the highest degree opposed to morality strikes everyone upon his thinking of it. Furthermore, the thought of it is so revolting that even calling such a vice by its proper name is considered a kind of immorality... However, it is not so easy to produce a rational demonstration of the inadmissability of that unnatural use....

Of course, the fact that few of us today share Kant's horror shows that there is no "hardwired" moral condemnation of masturbation that is "insensitive to environmental influence." But MFT assumes that nothing is hardwired or insensitive to influence. Rather, MFT posits that Kant, like the rest of us, had a domain-specific functionally specialized cognitive mechanism (the purity foundation) that attended preferentially to information about food, sex, and other bodily activities. It made it easy for Kant's society to teach children that masturbation is bad and to link masturbation to disgust during the course of child development. Even Kant was unable to think about morality by relying exclusively on his all-purpose undifferentiated domain-general intelligence, because Kant's mind was full of moral modules.

In conclusion, we are grateful to Suhler and Churchland for the extremely accurate overview of MFT that they offered in Section 2 of their essay and for the four points of praise that they offered at the end of that section. We believe that their three complaints in subsequent sections are not really valid complaints about MFT; two of them are better viewed as complaints by West Polers about nativist theories in general. MFT has been, from its inception, an attempt to bridge the nativism of evolutionary psychology with the constructivism of cultural psychology.

We freely admit that we built on sand. Morality is tough stuff to work with, and we are proud of ourselves for having solved the design challenges of doing so. Our house is not yet finished, and we welcome Suhler and Churchland's suggestions about where more work is needed.

Reprint requests should be sent to Jonathan Haidt, University of Virginia, or via e-mail: Haidt@Virginia.edu.

Note

1. See a list of publications by many authors reporting novel findings using MFT at www.MoralFoundations.org.

REFERENCES

Baron-Cohen, S. (2009). Autism: The empathizing-systemizing (E-S) theory. *The Year in Cognitive Neuroscience. Annals of the New York Academy of Science, 1156,* 68–80.

Barrett, H. C., & Kurzban, R. (2006). Modularity in cognition: Framing the debate. *Psychological Review, 113,* 628–647.

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior.* Cambridge, MA: Harvard University Press.

Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Fodor, J. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Graham, J., Haidt, J., & Nosek, B. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029–1046.

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., et al. (2008). Many sequence variants affecting diversity of adult human height. *Nature Genetics, 40,* 609–615.

Haidt, J. (2006). *The happiness hypothesis: Finding modern truth in ancient wisdom.* New York: Basic Books.

Haidt, J. (in press). *The righteous mind: Why good people are divided by politics and religion.* New York: Pantheon.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20,* 98–116.

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left-right: Ideological narratives and moral foundations. *Psychological Inquiry, 20,* 110–119.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus Fall, 133,* 55–66.

Haidt, J., & Joseph, C. (2007). The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3, pp. 367–391). New York: Oxford.

Herdt, G. (1981). *The Sambia: Ritual and gender in New Guinea.* New York: Holt, Rinehart and Winston.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and Emotion, 23,* 714–725.

Iyer, R., Koleva, S. P., Graham, J., Ditto, P. H., & Haidt, J. (submitted). Understanding Libertarian morality: The psychological roots of an individualist ideology.

Kant, I. (1996). *The metaphysics of morals* (M. Gregor, Trans.). Cambridge: Cambridge University Press. (Original work published in 1797).

Kass, L. R. (1997). The wisdom of repugnance. *The New Republic, June 2,* 17–26.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago: Rand McNally.

Koleva, S., Graham, J., Haidt, J., Iyer, R., & Ditto, P. (submitted). The ties that bind: How five moral concerns organize and explain political attitudes.

Marcus, G. (2004). *The birth of the mind.* New York: Basic.

Pinker, S. (2002). *The blank slate: The modern denial of human nature.* New York: Viking.

Samuelson, S. (1980). The cooties complex. *Western Folklore, 39,* 198–210.

Singer, P. (1979). *Practical ethics*. Cambridge: Cambridge University Press.
Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39–67). Cambridge: Cambridge University Press.
Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (pp. 53–68). New York: Oxford.
Suhler, C. L., & Churchland, P. (2011). Can innate, modular "foundations" explain morality? Challenges for Haidt's moral foundations theory. *Journal of Cognitive Neuroscience, 23,* 2103–2116.
Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science, 9,* 160–164.
Turkheimer, E. (in press). GWAS and EWAS.

This article has been cited by: