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.

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

University of Virginia

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.