

The Seeds of Humanity

MARC HAUSER

THE TANNER LECTURES ON HUMAN VALUES

Delivered at

Princeton University
November 12, 2008

MARC HAUSER is a professor of psychology and human evolutionary biology at Harvard University, where he is also director of the Cognitive Evolution Lab and codirector of the Mind, Brain, and Behavior Program. His research focuses on the evolutionary and developmental foundations of the human mind, with the specific goal of understanding which mental capacities are shared with other nonhuman primates and which are uniquely human. He has received numerous awards, including a Guggenheim Fellowship in 2005. He is the author of *Wild Minds: What Animals Really Think* (2000), *Moral Minds: How Nature Designed a Universal Sense of Right and Wrong* (2006), as well as the forthcoming volume *Evilicious: Why We Enjoy Being Bad*.

OVERALL INTRODUCTION

Humans create plays, operas, sculptures, computers, equations, laws, religions, guns, and soufflés. This is only a partial list of our achievements. In the history of life on earth, we are the only species to have created such creations. If a Martian landed on earth and had to develop a taxonomy of the living organisms, he could not be faulted for classifying the bees, birds, beavers, and baboons with the birches and baobabs, while placing humans in a group all on their own. After all, not only have baboons and baobabs never produced a soufflé, but they have never even contemplated the possibility. Baobabs lack the brains, whereas baboons lack the kind of brain that has both technological *savoir faire* and gastronomical creativity.

These observations suggest the first radical proposition I will make: we are not animals. Forget all the news about our shared genetic heritage with chimpanzees. If the fact that we share some 98 percent of our genes with chimpanzees is meaningful, then why isn't a chimpanzee writing this essay, or singing backup for the Rolling Stones, or working on quantum computing, or adjudicating over a legal case, or making me a soufflé? These facts about common genetic heritage just do not give us any traction into the problem of our uniqueness, our *humaniqueness*.

In addition to our exceptional achievements, we are also a paradoxically variable species. Different human cultures produce different languages, musical compositions, moral norms, and artifacts. From the viewpoint of one culture, another culture's expressions are often bizarre, sometimes distasteful, frequently incomprehensible, and occasionally immoral. Without doubt, the variation is massive, apparently limitless, and certainly meaningful to the individuals who share a particular culture. Nothing like this exists in other animals, even our closest relatives, the chimpanzees. Looked at in this way, a chimpanzee is a chimpanzee is a chimpanzee—a cultural nonstarter.

These observations lead, however, to a second radical proposition: the observed variation in human cultural expression, though unique, is superficial, concealing deep facts about our genetic constitution and the neural wiring it creates. What we perceive as differences between and within cultures is an illusion. The illusion is shattered, however, once we harness the discoveries of molecular biology and neuroscience to reveal four essential properties of human brain function:

- **Property 1:** The only way to generate limitless variation in expression is by means of recursive and combinatorial operations. Recursion is a looping operation, where a rule is called up over and over again, adding new expressions, be they longer sentences, new musical scores, or tools within tools (think Swiss army knife). The combinatorics allow us to combine and recombine discrete elements to create new representations.
- **Property 2:** Human creativity comes from our capacity to promiscuously play with thoughts from different domains of knowledge, allowing symbols for art, sex, space, causality, and friendship to combine, generating new laws, social relationships, and technologies.
- **Property 3:** We spontaneously convert analog representations to digital symbols, providing discrete elements for our recursive-combinatoric operations, and achieving great economy of computation.
- **Property 4:** Unlike animals, whose conceptual representations are anchored in sensory and perceptual experiences, many of our representational resources are highly abstract, with no clear connection to sensation and perception; language is one of many such systems.

Together, these properties provide a remarkable potential, but also a set of constraints that limit the range of *possible* languages, musical scores, moral rules, and technological devices. What creates specific cultural variants is a process of selection among the biologically presented options, building Korean or French, Bach or the Beatles, punishable or permissible killings, and spears or missiles.

To defend these two propositions, I develop four points in Lecture I. First, I provide a short preamble from research in evolutionary developmental biology (evodevo) showing that a core set of cellular operations and innovations, originating around the Cambrian some 500 million years ago, provided the source of all subsequent variation critical for constructing adaptive solutions to existing problems. On this view, much of the observed variation in animal anatomy and physiology, both extinct and extant, is superficial, relying on a basic blueprint, shared by all animals. Put simply, yes, there are tiny flies and massive whales, as well as spherical blowfish and cylindrical snakes, but there is one blueprint.

Second, I showcase the idea that 50 years of research in modern linguistics, initiated by the deep insights of Noam Chomsky, leads to a stunning conclusion: the variation observed among the world's languages is a *trompe l'oeil*, a dupe that makes us believe that we can create, willy-nilly, an unlimited variety of languages. Hiding beneath the surface of this canvas

is a set of universal computations, optimally designed to solve the problem of linking sound to meaning. Paralleling the work in biology, every human is born with a lingua kit—a *universal grammar*—for building a range of possible languages. Inside the kit are two essential elements: recursive operations and interfaces or connections between modules of the mind. Thus, to create any sentence, in any language, we have a recursive operation called *merge* that concatenates abstract symbols with syntactic structure (noun, verb, adverb, and so on), and then links up with our sound system to give voice to the thought that *It is an honor to deliver the Tanner Lectures* or *Is it an honor to deliver the Tanner Lectures?* but not *The is Tanner honor it lectures deliver*. Adopting this minimalist position not only clarifies a considerable amount of confusion in the study of linguistic structure but also helps solve a deep mystery about the evolution of language. Whereas previous inquiries stumbled over the apparent impossibility of a sudden emergence of language, with all of its complexity, only 150,000 to 50,000 years ago, the minimalist position helps solve this problem by pointing to the simple elegance of language. That is, the primary spark for evolving language, with all of its expressive forms, was the emergence of two basic ingredients: recursive operations and interfaces between modules of knowledge. The interfaces are novelties in the animal kingdom, and critical for language, as they link syntactic structures with concepts, and concepts with sounds, to create strings of words for rap music, theatrical prose, or stunning academic lectures.

Third, based on the parallel developments in biology and generative linguistics, and the fact that there was a sudden and surprising emergence of cave art, sophisticated cooking, musical instruments, complex weapons, and linguistic symbols, I suggest that this period in our history represents the starting point for our *cultured genome*. At this point, and not before, every healthy human was born with a capacity to create any language, music, moral norm, or artifact. This leads to two conclusions. One, the birth of new languages, musical expressions, moral institutions, and technologies should not trigger a celebration of our creativity. Rather, these novel expressions should inspire a toast to the evolution of a cultured genome some 100,000 years ago, a biological architecture that provided the *potential* to create such variation. Second, the variation is superficial, created by a core set of generative computations together with promiscuous interfaces between different domains of knowledge.

Fourth, by pursuing these ideas, and the currently available evidence, I provide an explanation for how, on the one hand, human and nonhuman animal mental life appears to share so many cognitive resources, and, on

the other, how human thought and its remarkable capacity for cultural expression seem unique—an explanation for our *humaniqueness*. Specifically, I suggest that humans evolved a distinctive cognitive architecture by building from a set of ancient mechanisms, adding a small number of unique generative processes together with interfaces between modular, domain-specific systems of knowledge. For example: though songbirds can riff like Charlie Parker on sax, combining and recombining notes to create new tunes, none of their variations adds new meaning in the same way that we humans, Parker included, can take this same combinatorial process and combine words to create new sentences; though animals have feelings, only humans cry with tears, both when sad and when happy, showcasing how our emotional system interfaces with our physiomotor system (our lacrimal glands); though animals can enumerate over food items or individuals, only humans integrate their linguistic system with the system of object recognition and morality to evaluate the ethical benefit of taking the life of one person to save the lives of many.

The road map ahead is as follows. I begin in Lecture I with a discussion of our mental uniqueness, and how these capacities triggered a cultural revolution. This section also discusses why the observed variation in cultural expression is illusory, based on a core set of generative capacities coupled with promiscuous interfaces between domains of knowledge.¹ In Lecture II, I take some of the ideas from Lecture I and apply them to a discussion of our moral sense, focusing in particular on the evolutionary and developmental origins of our intuitive ethics.² I then end with a brief discussion of how these ideas can be misinterpreted with respect to prescriptive claims about human nature and humanity, and how they can be properly interpreted and harnessed to create a more humanitarian space on the planet.

1. Many of the ideas and references in this section are discussed in a paper by M. Hauser: "The Possibility of Impossible Cultures," *Nature* 460 (2009): 190–96.

2. Several of the ideas discussed here are taken from a paper by B. Huebner, S. Dwyer, and M. Hauser, "The Role of Emotion in Moral Psychology," *Trends in Cognitive Sciences* 13, no. 1 (2009): 1–6; as well as recent books and papers (for example, M. Hauser, *Moral Minds* [New York: Ecco, 2006]; see also <http://www.wjh.harvard.edu/~mnkylab/publications/recent.htm>).

LECTURE I. HUMANIQUENESS AND THE ILLUSION OF CULTURAL VARIATION

1.1. The Cambrian Explosion and the Emergence of a Universal Genome

Nature presents us with a bewildering variety of animal forms, big and small, long and short, many legs and no legs, wings and gills, scales and feathers, and horns and tails.³ From Darwin's day through most of the twentieth century, there was little understanding of the source of such variation. The dominant view was that variation emerged as a result of random processes. Adaptive landscapes then followed, with peaks and troughs created by specific ecological and social pressures, favoring some forms over others.

Recently, new technologies and analyses have led to a different account of variation, inspired to a large extent by analyses of life on Earth during the Cambrian. Three observations are most relevant: (1) there was an unprecedented explosion of new life forms that appeared suddenly, en masse; (2) these animals, although anatomically and behaviorally diverse, showed remarkable similarity in their genetic makeup; and (3) though many of these animals were anatomically and behaviorally simple, they evolved what appeared to be an exuberance of genetic complexity. But, given the fact that such variation emerged all at once, and that simple organisms like worms and insects were equipped with genomes almost as large as our own, we emerge with one of the most exciting insights in the history of biology: much of the observed variation in animal anatomy and physiology, both extinct and extant, is superficial, relying on a basic blueprint shared by all animals. Variation emerges from this blueprint by means of a set of core cellular computations (rearrangement, repetition, amplification, and division; see fig. 1), together with compartmentalization of function and exploratory processes. If some of our Cambrian ancestors were alive today, they would look at this year's models and immediately detect a family resemblance of old and familiar parts.

To clarify the shift in focus, consider Darwin's classic example of adaptation—the Galapagos finches. On the classic view, we explain the variation in beak morphology by looking to details of their ecology, and, in particular, to how differences in seed size select for differences in beak

3. This section is presented in Lecture I, with comments by S. Gelman and H. Cronin.

Core cellular operations

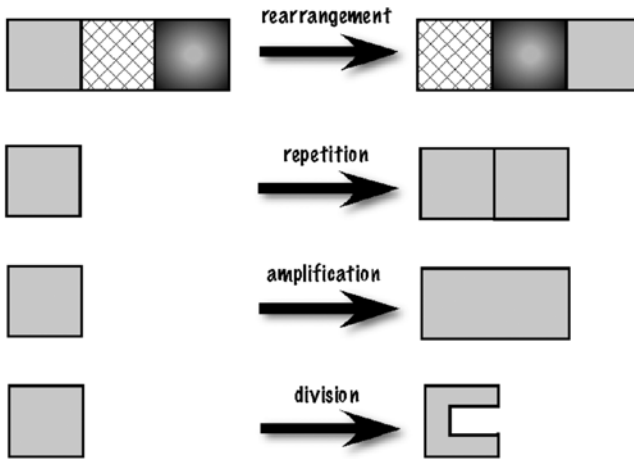


FIGURE 1. Core cellular mechanisms that enable variation in phenotypic expression without the introduction of new genetic material.

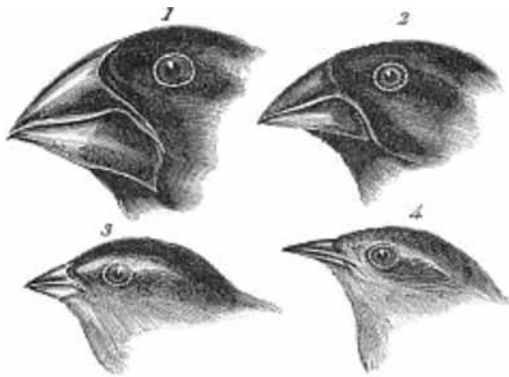


FIGURE 2. Variation in beak morphology among a sample of Darwin's finches.

shape, length, and heaviness (fig. 2). At one level, this is a satisfactory explanation of the observed variation, implying a guiding role for selection in coordinating anatomical changes with behavioral shifts in feeding ecology. Furthermore, these changes in beak anatomy had a direct effect on song structure, feeding directly back to mating restrictions and the role of sexual selection in guiding speciation.

Recently, evolutionary-developmental studies, led by Abzhanov, have provided additional explanatory power to Darwin's observations and the careful field research that followed. Early in ontogeny, as the beak develops, cells associated with the neural crest migrate away from the neural plate to a set of primordia around the mouth region. For large-beaked finches (fig. 2, pl. 1), the primordia express a protein associated with bone growth (Bmp) both earlier in development and at higher concentrations (note: amplification issue from fig. 1) than for smaller-beaked finches. With this information in hand, and the capacity to genetically alter closely related species, Abzhanov and colleagues experimentally inserted the Bmp protein into a chicken embryo. The result: a chick with a large, broad beak instead of its native hardware, the small beak. But this is not the most impressive result. More important, the chicken's new beak does not look like a poorly grafted accessory, created by an unskilled surgeon. Rather, the chicken's large beak develops seamlessly, much like that of its sister, the Galapagos large-beaked finch. What this result (and others like it) suggests is that selection need not coordinate the coevolution of anatomical components because developmental programs have been instantiated from the start to both create and facilitate such variation. To put it simply, the ancestral finch arrived on the Galapagos with the full tool kit for building beaks of varying shapes and weights, waiting for selection to pick among the options based on considerations of adaptive fit.

As a brief parenthesis to the issue of coevolution, it must be noted that even in cases where developmental programs provide the substrate for creating variation, this does not mean that neither the evolving nor the developing organism is devoid of coordination or interface problems. In particular, as new anatomical forms emerge, they must, at some level, integrate or interface with other systems to coordinate function. Once again, consider Darwin's finches. A large beak must be accommodated by musculature that can both support the weight of this beak and also maintain control of it to achieve functionality. Thus, bone morphology coevolves with musculature. Further, the morphology of the beak must be coordinated with the morphology of the wings to provide sufficient lift and stability during flight. As Abzhanov observes, when the large-beaked finch flies, there is a slight downward arch to the trajectory, a pattern created in part by the weight of the beak.⁴

This analysis of Darwin's finches, focused on one taxonomic group and one bit of anatomy, can be broadened to a wide variety of other problems

4. Personal communication.

in nature, covering a diversity of fauna and flora. For example, work by Bejan and colleagues over many years has demonstrated that the *designedness* that we see in nature (aerodynamics of raindrops, tree structures, animal locomotion) can be readily explained by considering the flow of energy from two positions in space, over time. Thus, it is not by chance that all treelike structures in nature—including river basins, blood vessels, lightning, and neurons—have, as their basic architecture, a scheme of binary branching. This configuration provides the optimal engineering solution when the requisite problem entails *moving* energy from one place to another. In a similar move, Bejan and Marden have shown that all forms of animal locomotion (running, swimming, flying) using a wide variety of body parts (legs, fins, wings) can be reduced to an optimization equation that links speed, body mass, and frequency. In particular, all forms of locomotion, in animals as different as running mammals, flying insects, and swimming crustaceans, are the product of a balance between the loss of energy in the vertical (lifting and then dropping the body) and horizontal (friction that arises from contact with the surrounding medium—land, water, air) dimensions. Last, and as discussed more completely below, studies of the skeletal morphology of the fauna from the Burgess Shale (at the midpoint of the Cambrian) reveal that a lean set of *seven* parameters accounts for not only all of the observed variation during this period but virtually all variation in skeletal morphology ever since.

Synthesizing, new methods and findings in molecular biology and morphology have inspired several biologists to argue that the observed variation in animal form is better explained by a theory of *facilitated variation* than by appeal to directed selection. And, further, the observed variation in a given environment or period of time reflects a process of *selection* among the biologically given options as opposed to *instructive* or *associationistic* creation and tuning of variation. This shift does not deny the role of natural selection in creating adaptations, but, rather, emphasizes the fact that the material to create such variation comes from other sources. In particular, the key idea is that weak regulatory linkage among a core set of cellular mechanisms, together with creative exploratory processes, accounts for the seemingly limitless variation we observe in animal form. As summarized by Gerhart and Kirschner, the primary architects of this theory:

Most anatomical and physiological traits that have evolved since the Cambrian are, we propose, the result of *regulatory* changes in the us-

age of various members of a large set of *conserved core components* that function in development and physiology. Genetic change of the DNA sequences for regulatory elements of DNA, RNAs, and proteins leads to heritable regulatory change, which *specifies new combinations of core components*, operating in *new amounts and states at new times and places* in the animal. These new configurations of components comprise *new traits*. . . . Of course the entire process is repeated in successive rounds of phenotypic variation and selection in an evolving trait.⁵

What is critical about weak linkage is that it enables core processes to be readily turned on or off, inhibiting or facilitating activity or expression. And what is critical about the exploratory processes is that they are both well designed and easily stabilized, properties that are essential when it is desirable to maintain randomly generated but functional variation.

Though biologists are certainly not in agreement concerning whether such processes are sufficient to account for the observed variation, what is striking about this characterization is its family resemblance to many concepts currently in vogue in the cognitive sciences and, especially, the study of language.

1.2. The Paleolithic Explosion and the Emergence of a Universal Grammar

Paralleling the observed variation in animal form, natural languages, extant and extinct, appear remarkably variable, including their sounds, lexicons, and organizational principles. Each language appears, in some sense, like a species with its unique anatomical form and courtship rituals. In the same way that the courtship displays of one species are unintelligible to another species, guaranteeing reproductive isolation, the sounds, meanings, and grammatical operations of one language make it unintelligible to a speaker of a different language.

Inspired by the work in cellular biology in the early 1960s, especially by Jacob and Monod, several linguists working in the generative tradition started challenging the significance and source of the observed variation in linguistic forms. In particular, the early work in linguistics suggested that the apparent variation was superficial, mediated by lawful regularities that, although impenetrable to introspection from most mature speakers of a language, could be uncovered by careful linguistic analysis. What this

5. J. Gerhart and M. Kirschner, "The Theory of Facilitated Variation," *Proceedings of the National Academy of Sciences* 104 (2007): 8582–83 (emphasis added).

line of inquiry soon revealed was a set of universal computations that, together with dedicated systems for conceptual and phonological representation, provided a family of developmental options for building different languages. Though there are controversies concerning the limits of linguistic variation, and the computations required to account for such variation, here I briefly sketch the logic of the generative tradition, focusing specifically on parallel theoretical distinctions and findings from evolutionary developmental biology.

A first point of contact is the fact that children are born with the capacity to acquire a wide range of *possible* languages, as opposed to specific languages such as English, Korean, or French. What this implies is that the child is equipped with an acquisition device that must be abstract, enabling the growth of many different languages, each with its specific sound structures, lexicon, and rules for arranging these items. Further, by considering the idea that the child's acquisition device generates a space of possible languages, it immediately becomes necessary to consider the idea that something internal or external to the device creates a space of *impossible* languages, forms that are never even entertained by the child because they are poorly designed for acquisition and externalization. This proposal maps on to Peirce's abductive principle (a point first noted by Chomsky), whereby only a restricted set of hypotheses is even considered in the context of a set of data; such constrained searches are also of relevance to Bayesian analyses that attempt to account for patterns of acquisition, including the role of negative evidence.

The beauty of thinking about the child's linguistic endowment as a system for building a space of languages is that it maps onto a considerable amount of work in functional morphology exploring the space of possible forms. Thus, in the same way that biologists speak of *morphospaces*, I suggest that it is not only reasonable but necessary to speak of *linguaspaces*. Let me flesh out the logic of this claim by considering research on the design space of skeletal morphology. Based on detailed descriptive work at the functional morphological level, analyses reveal a library of seven states or parameters that account for virtually all of the variation observed since the middle Cambrian, specifically the appearance of the Burgess Shale fauna (fig. 3).⁶ Each variable has a set of options, set by both endogenous and exogenous factors that arise over evolution and in development.

In parallel, research in generative linguistics suggests that a core set of operations, some universal (for example, principles) and others optional

6. R. D. K. Thomas, R. M. Shearman, and G. W. Stewart, "Evolutionary Exploitation of Design Options by the First Animals with Hard Skeletons," *Science* 288 (2000): 1239–42.

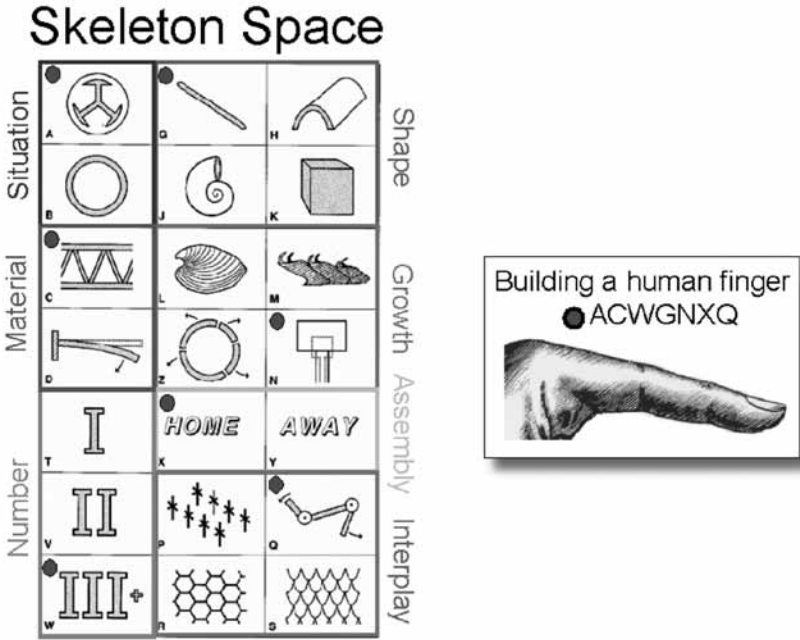


FIGURE 3. The skeletal space of Thomas et al. that yields potential forms. There are seven core properties (colored boxes), each with two to four possible states (cells within boxes), for a total of twenty-one variables: **Situation (Location of skeleton)**: A, internal; B, external; **Material composition of elements**: C, rigid; D, flexible; **Number of elements**: T, one; V, two; W, three or more; **Shape of elements**: G, rods; H, plates; J, cones; K, solids; **Growth of elements**: L, accretionary; M, serial units and branching; Z, replacement/molting; N, remodeling; **Assembly of elements**: X, growth in place; Y, prefabrication; **Interplay of elements**: P, no contact; Q, jointed; R, sutured or fused; S, imbricate (that is, folded over, overlapping). A human finger is designed on the basis of one state (red circle) from each of the seven properties, specifically, ACWGNXQ.

(for example, parameters), provides not only the source of our capacity for linguistic expression but a system of constraints on the linguaspace. Thus, all languages rely on operations such as recursion, copying, movement, and displacement, generating hierarchical structures. Many of these operations are constrained, however, by aspects of computational efficiency, as well as by domain-general considerations such as memory and learning. Paralleling work on morphospaces, we can imagine two n-dimensional structures or linguaspaces (fig. 4). One represents a purely theoretical construct based on a set of presumed parameters that are necessary to construct language. A second, similar in kind, represents a more empirically anchored space, using parameters that are known to be relevant to

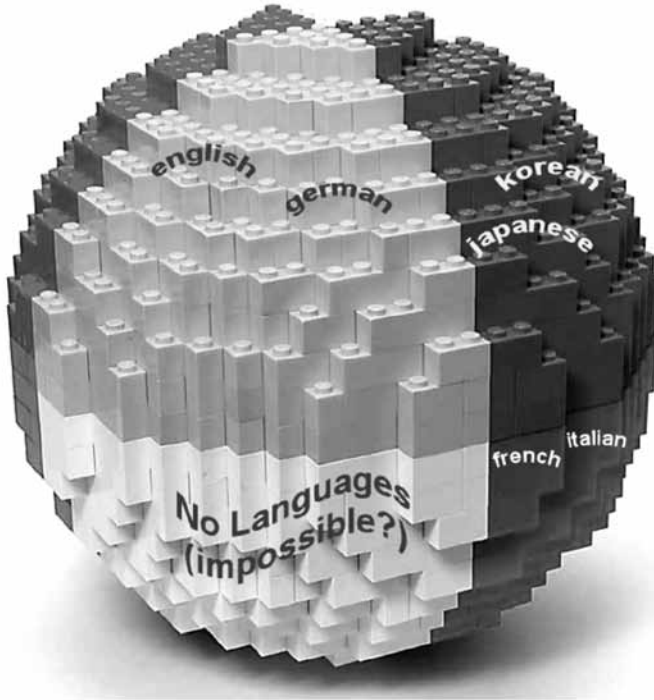


FIGURE 4. A linguaspace is an n -dimensional structure that represents possible languages. Here, only a few languages and their nearby relatives are marked. The area demarcated in white is a part of the space that lacks representation from extant or extinct languages, and, thus, may capture the zone of impossible languages. Missing, of course, are the dimensions that would define this space.

actual natural languages and with regions of the space filled by such languages. In both cases, there is an opportunity to define the range not only of possible but of impossible languages, anchored in a region of the linguaspace. For the second, empirically defined, linguaspace, empty regions can be explored, attempting to understand the nature of constraints—that is, why there are no extant or extinct languages in this space.

What makes this enterprise challenging, of course, is defining the N —that is, the dimensions that provide structure to the space. Although this challenge confronts theoretical morphologists as well, it is greatly exacerbated in the study of language. That said, studies of artificial language learning, computer modeling, and theoretical linguistics all hold out hope of reducing the uncertainties.

A second point of contact between evodevo and linguistics concerns the nature of the input and the timing of growth and development. When

a child grows its native language, the appearance of certain structures is constrained by the appearance of others, as well as by the timing and magnitude of the input. Some of these constraints are specific to language, and some result from the interaction between language-specific operations and processes that are more domain-general. For example, though recursive operations such as *merge* (loosely, an iterative operation that takes two elements and combines them into a set to create new expressions) are unlimited with respect to the number of iterated computations, they are constrained by language-external processes of memory and comprehension, as well as by properties of the motor system that enable externalization, forced through a process of linearization. By analogy, much of the work in *evodevo* suggests that the growth and development of different animal forms arise due to core operations such as rearrangement, repetition, magnification, and division, with each of these processes further modified by the timing and magnitude of experience, as well as the locus of change.

A third and final point of contact concerns how the internal language system ultimately forms an acquired and externalizable language. If, as discussed above, the acquisition device is a generator of possible languages—the *linguaspaces*—then the role of the environment is to *select* among the possible options. This selective view of language acquisition is set in opposition to an instructional process in which the information or knowledge comes from the environment. Put simply, the distinction is between the environment as *critical shopper*, selecting options, as opposed to the environment as *pedagogue*, providing the options. I suggest here that the critical-shopper metaphor, and selection more specifically, more accurately captures the process of creating the observed cultural variation, especially given the evidence for this process in other natural systems, including the immune system, the development of animal forms, neuronal wiring, and the acquisition of birdsong. For example, songbirds have evolved brains that use a finite set of note types to create a highly variable range of possible songs. Depending on the environment, certain note types are selected and reproduced in particular orders. Similarly, the immune system generates a massive range of potential responses, but it uses input from the environment to lock onto a particular immune response. And, last, the synthetic writings of Changeux and Edelman all point to the idea that the brain generates an exuberance of potential connections in development, only some of which survive to become functional as a result of selection from the available options.

In sum, research in the generative tradition of linguistics suggests that, like the variety of animal forms, the variety of languages is superficial, concealing lawful regularities with respect to the underlying mechanisms for generation and acquisition, constraints on the space of possible languages, and a critical link between the core computations and interfaces both within the linguistic system (syntax, semantics, phonology) and outside of it.

1.3. Generativity, Promiscuity, and Illusory Cultural Variation

In section 1.1, I reviewed the molecular evidence indicating that much of the observed variation in animal form is due to core cellular processes that are generative, together with two key cellular innovations: weak regulatory linkage among core, modular functions together with creative exploratory processes. In section 1.2, I suggested that the generative tradition in linguistics licenses the conclusion that there were three essential innovations in the evolution of language: generative computations, interfaces between core modular components (syntax, semantics, phonology), and creative exploratory links into other domains of knowledge. Here, in this section, I suggest that by combining the insights from these two disparate traditions, we alight upon an intriguing road map into the territory of brain and cognitive function, the pieces that ultimately distinguished *Homo sapiens* from the other primates. Specifically, I suggest that over human evolution there was a change from highly modular systems with few interfaces to weaker modules with numerous, promiscuous, and combinatorially creative interfaces. This system generates, with lawful regularity, cultural options. The environment, qua *critical shopper*, then selects among the options, ultimately stabilizing the system to yield a specific cultural variant. This process of generate, test, and stabilize is the signature of our cellular machinery!

Consider the creation and diversity of human artifacts—especially tools—in a comparative, evolutionary, and developmental context. Unlike (even) our simplest tools, such as the pencil (fig. 5), animal tools are created from one material, for one function, are most often dispensed after their first usage, and are never used for functions other than the original one. The first two features reveal that, unlike human tools, the representation of animal tools lacks the combinatorics (for further discussion, see section 1.4). A pencil combines four materials (lead, wood, metal, rubber), to create four functions (lead for writing, wood for holding the lead, metal for holding the rubber to the wood, rubber for erasing). Moreover, each



FIGURE 5. How our combinatoric brain generates a pencil:
lead + wood + metal + rubber = writing and erasing tool.

material can be used for a variety of other functions, including rubber's role in chewing gum. As experiments reveal, ask a young child what she can do with a pencil other than write, and she will immediately offer such functions as holding up her hair, puncturing a plastic cover, and poking a friend. Only our species thinks of artifacts as designed for a particular function. However, due to promiscuous interfaces, we can also think of many other functions. More generally, our minds generate numerous design options, allowing particular environmental experiences not only to signal a specific design but also to stabilize this design within a culture, at least until new demands put pressure on change.

Consider, again, the pencil. If the material to be written on is wet, this environmental condition signals to the writer that a pencil will not work. In contrast, a dry material, with some absorptive texture, satisfies the function, and stabilizes the pencil as a tool with good design specs. But in the case of tools, and tool users like us, the immediate environment is not the only consideration for stabilization. There are also future environments or considerations; for example, if the written material is intended to have legs, lasting longer than the average sticky-note message, then the environment sends a signal to destabilize the functional effectiveness of a pencil, favoring other materials for inscription.

One way to connect the discussion of artifact design back to the design of animal form is to consider the role of segmentation, and one of the



FIGURE 6. Recursive iteration of segments in centipedes and telescopic cups

core generative computations, recursion. This is the capacity that allows us to iteratively call up a rule to generate limitless variation of expressive forms. Now consider the centipede and telescopic cup (fig. 6). Here we see the invention of segmentation, a design feature that emerged during the Devonian, at a time when we first see annelids (worms) and arthropods (for example, insects, crustaceans); in cultural evolution, segmentation is harder to date due to the fragility of the archaeological record, but appears during the mid-Pleistocene with the presence of spears combining metal with wood. Once segmentation emerged, it was relatively trivial to copy such units many times until constraints of physical design intervened to end the iteration. The recursive operation is simply an iteration of: *join another segment*. This recursive rule stops generating new forms when the design—here, the number of segments—no longer functions to solve the problem. In the same way that the environment selected among the segmental options for centipedes to achieve some balance between size and locomotory efficiency, the environment of R&D selected among the options for telescopic cups to achieve some balance between stability and a compacting container for liquids. Again, this kind of analysis should constrain the pride we place upon our technological creativity. Instead, we should stand in awe at the idea that the genetic revolution that took place some 150,000 to 50,000 years ago set us up with a genetic recipe for creating an exuberance of technologies, or, to paraphrase Darwin's words, endless artifactual forms most beautiful.

This way of looking at artifacts also enables a strong connection to Biederman's research on object perception (fig. 7) and, in particular, the idea that one can break down the recognition process into components called *geons*—volumetric units—that, when combined, provide the essential elements for recognizing and, perhaps by extension, building objects. And intriguingly, many of the parameters that establish Biederman's geon-

Object Construction Space







Geon	Edge	Symmetry	Size	Axis
	S	++	++	+
	C	++	++	+
	S	+	-	+
	S	++	+	-
	C	++	-	+
	S	+	+	+

FIGURE 7. A partial set of Biederman's geonspace (volumetric primitives) for object recognition by components. Edge has two options: straight (S) or curved (C). Symmetry has three options: symmetrical under rotation at 90 degrees and reflection (++), reflection only (+), and asymmetrical (-). Size has three options: Constant (++), Expanded (-), and Constant and Expanded (--). Axis has two options: Straight (+) and Curved (-).

space parallel those built into Thomas and colleagues' work on the skeletal morphospace, including concerns for symmetry, size, and the axis for rotation and stabilization.

The combination of weakened modularity and enhanced interfaces that facilitates the observed cultural diversity has two additional consequences. First, whereas other animals typically generate narrow solutions to problems—exhibiting a form of *myopic* intelligence—humans explore alternative solutions to the same problem and often extend a solution from one context to another. Second, because we are endowed with promiscuous interfaces, we readily take the ancient systems of knowledge that we acquired from our primate relatives and transform them into new systems of thought. Let me illustrate each of these points.

Over the past thirty years, we have learned an extraordinary amount about animal behavior, including the adaptive logic of their actions and the psychological instincts that guide their decisions. One thing is clear: many of the capacities that we once thought were unique to humans are clearly shared, in some way, with other animals. That said, the operative phrase in the last sentence is “in some way.”

As figure 8 highlights, animals reveal many capacities that have some family resemblance with humans. But when these capacities are explored in detail, a striking difference emerges that maps onto the first point above: the solutions that animals have evolved are local, solving only the specific adaptive challenge presented. Thus, for example, von Frisch considered the honeybee’s dance as a “language” because it exhibited some of its core properties: for example, the communicative signal is, in some ways, detached from the context—foragers go out and find food, and then come back to the hive to communicate to others where it is; the dance is not an emotive response but seems to represent in some way abstract information about distance, direction, and quality of food. Though these certainly are properties of human language, if we want to tell a complete evolutionary story, we need to understand in what way the mechanisms underlying these properties are similar to or different from those underlying human language. And here is where the fundamental difference emerges: honeybees do not use this capacity, whatever is, in any other context but foraging. One possible reason for this contextual myopia could be because foraging is really all there is to bee life. In fact, nothing could be further from the truth. Bees, like other eusocial insects, have remarkably rich social lives, with division of labor, policing, cheating, and intricate house-keeping. Alas, the bees never seem to really communicate about these events, and insofar as they do, the signals deployed are nothing like their dance, and, thus, nothing like human language. What we can say, therefore, is that whatever this representational system is for the honeybee, it lacks the arbitrary relation between meaning and externalized signal that makes human linguistic expression possible, and virtually unlimited. The same distinction can be applied to all other *nonhuman* animals.

A second capacity that has long been hailed as uniquely human is teaching. As many have discussed, teaching plays a pivotal role in cultural evolution. It is the cheapest mechanism we have for disseminating information to large numbers of individuals, in one shot. By giving lectures, I am able to implant an idea in the heads of hundreds of audience members



Ants that know the location of food lead naive individuals to the food source, showing sensitivity to their movements that indicates a form of teaching. This kind of teaching is, however, restricted to this particular food-finding context. Source: N. R. Franks and T. Richardson, "Teaching in Tandem-Running Ants," Nature 439 (2006): 153.



Jays not only recall what kind of food they have stashed, but where they stashed it and when. The time stamp of when is critical to what is often described as episodic memory. In jays, however, the capacity appears restricted to the context of food-cache recovery. Source: N. S. Clayton and A. Dickinson, "Episodic-Like Memory During Cache Recovery by Scrub Jays." Nature 395 (1998): 272-74.



Meerkats show evidence of teaching their young how to hunt dangerous scorpions. This capacity, which reduces the risk of danger to the pups, and helps them improve their prey-capture skills, is not extended to any other context. Source: A. Thornton and K. McAuliffe, "Teaching in Wild Meerkats," Science 313 (2006): 227-29.



Plovers engage in an injury-feigning display that lures potential predators away from the eggs in their nest. The display is deceptive as the plover is not injured, and if the predator persists, the intensity of the display continues until the predator moves away from the nest. Such deception is not carried out in any other context. Source: C. Ristau, "Aspects of the Cognitive Ethology of an Injury-Feigning Bird, the Piping Plover," in Cognitive Ethology: The Minds of Other Animals, 91-126 (Hillsdale : Erlbaum, 1991).



Chimpanzees, in a competitive context, can use the direction of another's eye gaze to draw inferences about what that individual can see, and thus, what its likely goal is. Chimpanzees fail to use eye gaze, however, in more cooperative contexts. Source: B. Hare and M. Tomasello, "Chimpanzees are More Skilful in Competitive than in Cooperative Cognitive Tasks," Animal Behaviour 68 (2004): 571-81.

FIGURE 8. The myopia of animal intelligence. Each of these examples illustrates that animals evolve adaptive solutions to narrowly defined problems, lacking the kind of generality seen in human cognition.

sitting in a lecture hall, and by putting my lecture on the Internet, I am able to infect thousands. And if you are Michael Jordan giving a tip about basketball, or Bill Gates offering a thought about investments, you can infect millions with a short YouTube message.

Recently, studies have turned up evidence of teaching in two unexpected species: ants and meerkats. In the case of ants, Franks and his colleagues observe that an individual who knows where food is located (the teacher) moves with a naive individual (the pupil) in what Frank and his colleagues call *tandem running*, a choreography in which the teacher constantly checks the pupil's movements, going back for guidance when they are off course. This form of instruction is clearly costly to the teacher who could merely walk on to the food source and eat. In meerkats, Thornton and McAuliffe have observed hunter-savvy adults provide their naive pups with opportunities to kill deadly scorpion prey, modifying the nature of the opportunity as a function of the pup's age. Though these are fascinating behaviors, and certainly function to provide naive individuals with relevant experiences, such structured instruction never arises in any other context. Again, one might imagine that these are the only contexts where teaching is necessary or relevant, but that is hard to believe. Ants are eusocial insects like bees, and thus, given the complexity of their society, why not engage in a little bit of pedagogy with respect to the slackers in the community, or have the soldiers inform everyone about impending threats? And for meerkats, whose stock has risen thanks to the wisdom of an agent who put them on the smash-hit television program *Meerkat Manor*, we know that they too live in complex social groups, competing and cooperating. Presumably, there is much that an adult could teach its young about who is tough and who is soft, who is hot and who is not. Alas, no teaching occurs outside of hunting for scorpions.

This story, of narrowly targeted, adaptive solutions, runs throughout the animal kingdom. It shows, I believe, that animal brains are hypermodular, encapsulated devices that evolved to solve one problem and solve it well.

The second point concerns the role of interfaces. To explain how an interface works, consider the now extensive research on numerical representations and computations in nonhuman animals. Dozens of studies indicate that even though human infants and nonhuman animals lack language, they nonetheless have the capacity to quantify objects and events in the environment, both with and without training from human experimenters. One system computes over analog magnitudes and provides an approximate calculation, limited by Weber ratios. A second system, often referred to as the process of parallel individuation, computes over distinct individuals and provides a precise calculation, limited to numbers less than about four. A third system, only recently described, computes

over sets, with successful discrimination of one from many, but not many from many. This set-based quantificational system looks like a potential precursor, conceptually at least, to the linguistic distinction, seen in many languages, between singular and plural. That is, when we express the plural form of the word “cup” as in “cups,” we use this marked form whether we have 2,100, or 5,000,000 “cups.” At this level, the rhesus monkey system looks the same as the linguistic system of singular and plural. That is, it distinguishes singular/one from plural/many, but not plural/many from plural/many. But this parallel fails to push deeply enough into the morphosyntax of singular-plural. To make this clear, fill in the blank with either “cup” or “cups” for the following quantities 0, .5, -3, and 1.0 _____. If you are like most English speakers, you will have used “cups.” The reason is simple: the marked form uses “cups” for anything that is not 1, and only 1, including the numerically equivalent 1.0. Thus, the abstract principle of morphosyntax transforms the representation—the one shared with primates—into something completely different.

Another example of interfaces begins with the discovery of mirror neurons, cells originally discovered by Rizzolatti and colleagues in the premotor cortex of macaque monkeys. These cells fire when the individual either produces an action or perceives an action. For example, some cells respond when the animal either grasps or sees another individual grasp an apple, other cells fire to the sound of a peanut cracking and to the sight of seeing someone crack open a peanut, and yet others fire to the perception or production of more abstract goal-directed action as evidenced by the fact that they fire to a hand grasping food or a mouth contacting the same food. In general, then, these cells provide an important substrate for action comprehension, anchoring at least some aspects of understanding of what others do in what the perceiver *can* do.

These beautiful physiological results in macaque monkeys led to the discovery in humans of potentially homologous regions of activity during neuroimaging experiments. Although imaging studies provide only population-level activity, results revealed significant activity in the premotor cortex (and other regions such as the insula) with the signature of mirror neurons: that is, coactivation during perception or production of action. The interface issue emerges when we consider the variety of contexts in which these areas activate in humans. In particular, several studies show that areas such as the premotor cortex and insula activate during production- or perception-based tasks involving imitation, empathy, and mind reading, with some authors suggesting that the breakdown of mind

reading in autistics reflects the breakdown of the mirror-neuron system. Although there is controversy concerning the extent to which mirror neurons represent the workhorses for such psychological processes, it seems clear that they are involved, contributing to the intersubjectivity that underpins imitation, empathy, and mind reading. Importantly, however, evidence for imitation in macaques is virtually nonexistent,⁷ there is no evidence of empathy, and mind-reading capacities appear limited. What these comparative results suggest is that the mirror-neuron system is not sufficient for imitation, empathy, and mind reading. Over the course of human evolution, this system either replicated to form connections with other neural components, or its connectivity changed to provide interfaces to these other components. Thus, and as argued by Rizzolatti, there are mirroring systems for action and for emotions, and these have different functions or goals. And once these new interfaces emerged, it was possible to extend our intersubjective world to cover the execution and interpretation of actions and emotions in a diversity of contexts, recruiting spatial, social, and moral knowledge.

In addition to the adaptive and evolutionary consequences of thinking about cultural variation in terms of promiscuous interfaces among modular systems of knowledge, there are two others, one conceptual and one methodological. Conceptually, this perspective makes human creativity and imagination seem much less impressive, and human nature (read: our biology!) much more impressive. That is, we were handed, by evolution, a tool kit for creating cultural variation in linguistic, musical, artifactual, and moral expression. This tool kit consists of a suite of developmental programs that generate variation, the raw material for a selective process that crystallizes to a particular form of expression. Methodologically, this perspective points to a weakness in many accounts of cross-cultural variation that rely on cataloging past and current cultures. In particular, if we are equipped with developmental programs that can generate a space of cultural expressions, the observable cultures may occupy only a small fragment of the potential space. To uncover whether the currently empty space is within the range of theoretically possible cultures, we must supplement our descriptive observations with experiments, assessing which cultural variants are intelligible, acceptable, and learnable. Of those that fail to

7. Although there is one report by Subbiaul and colleagues for a form of motor imitation, and a second report by Ferrari and colleagues of early infant facial imitation, these two cases are isolated, set in the context of many failed attempts to observe imitation in macaques or other Old World monkeys. Further, even if these two cases replicate, they fly in the face of no evidence of social traditions in macaques, and no evidence of vocal dialects, two situations in which one would expect to see the signature of an imitative species such as our own.

meet these design criteria, we may probe further in an attempt to understand the nature of the constraints.

Some of the ideas above have already begun to proliferate outside of linguistics into the less well-studied domains of music and morality, with intriguing experimental evidence and novel theoretical insights. For example, Jackendoff and Lerdahl have drawn an explicit analogy between music and language, using some of the core concepts from generative linguistics to explore the universal principles of organization underlying tonal compositions. Both music and language share core resources, such as the use of combinatorial operations and the representation of hierarchical structure, analyses that have been supported by neuroimaging studies revealing common brain regions. Music and language also operate over different representations, and tap different operations, as revealed by patient populations showing dissociations. Further, the computations that appear specific to music interface with more domain-general processes such as emotion, to generate musical preferences, and methods to manipulate them. Last, as Lerdahl's book title reveals, an important aim of this work is to map out the tonal music space—that is, the range of possible musical forms, together with the mechanisms that generate and constrain such forms.⁸

Similarly, and in an even more immature state of development, Harman, Dwyer, Mikhail, and I have all made an explicit analogy between language and morality, building off an insight first made explicit by Rawls. Here as well, evidence suggests that like language, some of the computations underlying our moral judgments operate outside of our awareness, are abstract, are highly generative (combinatorics operating over actions to create new, meaningful events), and show considerable immunity to cross-cultural variation. Like music and language, the moral domain also gains its generative power by means of novel interfaces between core domains of knowledge. Thus, for example, we typically perceive a difference between actions and omissions (all else equal), in part because actions enable a more transparent reading of causal responsibility and intentionality than do omissions. Take the outputs of this domain-general distinction and the systems that handle analyses of consequences to individual welfare, and we generate the judgment: *actions are worse than omissions*.⁹ Last, though we

8. F. Lerdahl, *Tonal Pitch Space* (New York: Oxford University Press, 2001).

9. By domain-general I do not mean that this distinction is necessarily unrestricted, applied to any situation in the way that our systems of memory or attention are domain general. It is certainly the case that intentionality represents a folk psychological notion, whereas cause applies in the case of both psychological and physical agents.

certainly do not have an understanding of the *possible* moral systems—the moral space—charting its design emerges as an inevitable outcome of the perspective taken here. I pick up this problem in Lecture II.

What studies of music and morality suggest, then, is that, like language, there is a core set of computations, combined with a creative set of interfaces between different modular systems, that generate massive variation in the range of potential musical compositions and moral norms. Although language, music, and morality all use some of the same computations (for example, combinatoric operations), it is not yet clear whether the brain houses these for all systems to use, or whether each domain has its own separate but identical set of computational resources. Further, what appears to give each domain its unique *feel* as a domain is how different components or modules interface to create novel representations. Thus, for example, both music and language rely on hierarchical structures, but language operates over lexical items that are ordered syntactically, whereas music operates over notes with specific durations and frequencies that are ordered over time. Similarly, though both morality and language tap some of the same conceptual resources, language maps these to phonology to create words, whereas morality maps concepts to actions to create morally meaningful (interpretable) events. Though we clearly hear the difference between Bach, Beethoven, Bartok, and the Beatles, musical genres from different eras and cultures, the differences are based on a common set of computational operations, linking temporal and spectral patterns of sound that ultimately interface with our emotions to stir either a pleasant or an aversive feeling.

1.4. Humaniqueness

I end this part of the essay by rethinking what it means to be human. I start by laying out six characterizations of human thought—why it is as we observe it to be rather than some other way—that I assume are relatively well accepted and, thus, uncontroversial:

1. There are human universals characterizing both our inner mental lives and our capacity for expression—all humans experience a core set of emotions, symbolize thoughts, imagine what others think and feel, play music, express their ideas in a particular language, and create artifacts.
2. There is variation in how humans express themselves as a means of signaling cultural identity—we play and enjoy different kinds of music, speak different languages, develop different rules to punish vice and reward virtue.

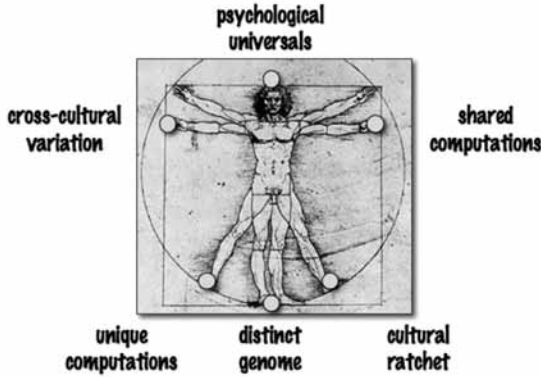


FIGURE 9. The essence of human thought and cultural expression.

3. We share several cognitive abilities with other animals—essential aspects of numerical quantification, spatial representation, tool use, music perception, pattern recognition and categorization, social knowledge, and economic decision making are seen in a variety of nonhuman animals, both closely and distantly related.
4. We have several uniquely human cognitive abilities (fig. 9): recursive and combinatorial thought, spontaneous symbolic production, promiscuous interfaces between different modules or domains of thought, and abstract conceptual systems that are detached from sensory and perceptual experience.
5. Given our unique cognitive abilities, something about our genetic constitution made possible this evolutionary change.
6. The ratcheting effect of cultural, as opposed to biological, evolution has greatly increased the gap between us and them.

Given these six characterizations—a taxonomist’s description of our species-specific anatomy—we arrive at an explanation for what makes us distinctively smart, adaptable, and virtually unlimited in our capacity for cultural expressions, while also explaining why our creativity, and the cultural variation it generates, may be less impressive than we imagine, built, as it is, out of a core set of conserved cognitive operations. The universality grows out of both an evolutionarily shared cognitive architecture, together with a small number of unique psychological processes—traits 3, 4, and 5. These processes constrain what humans can do, but provide an abstract, generative tool kit for building possible languages, musical compositions, moral norms, and technologies—in the lingo of theoretical morphology,

we have linguaspaces, musicospaces, moralspaces, and technospaces. The extensive cultural variation is possible only because of the universal core. That is, we uniquely evolved a set of generative processes together with promiscuous interfaces that were only weakly regulated or controlled, enabling different domains of knowledge to productively combine, yielding new solutions to fundamental problems posed by novel environments. A minimalist summary of the unique core of the human faculty reduces to Recursive Thought *plus* Interfaces. By creating interfaces, the brain generates novel cultural expressions, some that survive and others that are selected out in a process that may mimic more basic biological processes such as the immune system; again, the environment is more productively construed as a picky shopper than as a sagacious pedagogue.

Consider the invention of controlled fire. Though we do not know how this capacity emerged, its expression and subsequent use represent the combination of several different faculties, including folk physics (fire changes the physical properties of objects), folk psychology (fire changes the internal states of humans, making them warmer), and causality (the effects of fire on an object are irreversible). As some have argued, with the invention of controlled fire, the human potential was transformed, allowing our species to uniquely invade new and previously inhospitable environments, surviving in severe cold and eating previously inedible foods.

I very much doubt we could even teach another animal to make a fire, and even if we could, I doubt that they would see the power of its uses. The same comment could be made about numerous other capacities that represent the outcome of promiscuous interfaces, or to quote Sir Lawrence Olivier commenting on Marilyn Monroe: "Teaching her to act [is] like teaching Urdu to a marmoset." But teaching other humans and allowing them to imitate represent the most powerful vehicles for cultural diffusion. Once the information is disseminated, however, humans have a choice to accept or reject what is on offer. It is this freedom to choose, and to understand that there are alternatives, that is liberating when it works, and exceedingly painful when the options are restricted or, worse, taken away. And it is at this stage where our moral sense clicks in, and we hope, a broader community of ethically concerned citizens.

LECTURE II.
TO DO, OR NOT TO DO —
THAT IS THE MORAL QUESTION

In the Cohen brothers' recent blockbuster *No Country for Old Men*, Javier Bardem plays Anton, a psychopath who is smart, rational, and manipulative.¹⁰ What makes him a psychopath, as opposed to some other run-of-the-mill killer, is his dispassionate approach to others. When he kills, it is effortless, cold, disconnected, liberated from the emotional constraints that guide most of our actions and subsequently feed back on our actions to remind us of the good and bad, and to hand us moral appreciation. On this reading, Anton looks like he has lost his moral compass. But there is also a reoccurring scene in the movie that pushes a different kind of interpretation. As Anton searches for the man (Llewelyn) he believes has stolen his money, he meets several people and offers them an opportunity to “call it,” that is, to say which way a coin will land after being flipped—heads or tails. The following scene, between Anton and Carla (Llewelyn's wife), captures the essence of his psychology:

CARLA: You don't have to do this.

ANTON: (*smiles*) People always say the same thing.

CARLA: What do they say?

ANTON: They say, “You don't have to do this.”

CARLA: You don't.

ANTON: Okay.

(*Anton flips a coin and covers it with his hand*)

ANTON: This is the best I can do. Call it.

CARLA: I knowed you was crazy when I saw you sitting there. I knowed exactly what was in store for me.

ANTON: Call it.

CARLA: No. I ain't gonna call it.

ANTON: Call it.

CARLA: The coin don't have no say. It's just you.

ANTON: Well, I got here the same way the coin did.¹¹

There are two lines that point to a different interpretation of the psychopathic mind. First, when Anton responds to Carla's statement that he

10. This section is presented in Lecture II, with comments from S. Blackburn and W. Sinnott-Armstrong.

11. See <http://www.imdb.com/title/tt0477348/quotes>.

does not *have to do it* (that is, kill her), he responds that the “best he can do” is offer her a coin toss. This suggests a limited understanding of options, or a clear sense of options but only a limited desire to sample from the set. Second, and what is most telling, is Anton’s response to Carla’s comment that a coin cannot have a say in the decision to end her life, that only he, Anton, gets a say. Anton then says that he is dictated by chance, just as the coin is so dictated. Here we have a different view of the psychopathic mind, one that is predetermined, and in some sense lacking in free will. In the same way that a coin toss is simply affected by the physics of force and gravity, Anton’s reply implies that he is similarly guided by forces outside of his control. The question is: what are the forces that guide his decisions and the actions that flow from them? What is the right descriptive characterization of how Anton decides what to do, and does he even entertain what he ought and ought not to do? These questions take us to the heart of Lecture II: what is the most accurate way to describe the principles or processes that underlie our judgments of right and wrong, to what extent do these factors (the *is* of our moral capacity) influence our conceptions of what would be best for society (the *ought* of our moral capacity), and what happens when neuropsychological problems emerge that compromise the processes that underlie the *is* or the *ought*, or both?

I develop the argument as follows. First, I briefly sketch the theoretical landscape of ideas that have been brought to bear on these issues. Since much of this landscape has been reviewed elsewhere, I will focus the discussion on how different theoretical perspectives make predictions that the empirical sciences can, and recently have, tested. The primary distinction here will be between controlled and automatic processes, on the one hand, and then, within the automatic processes, the significance of emotions in guiding moral judgments as well as morally relevant actions. Second, based on the theoretical issues raised, I tackle head-on the evidence that has been mounted in favor of the sentimental perspective—that is, the Humean position that emotions provide the source of our moral judgments. I conclude that the current evidence does not support the hypothesis that emotions are causally prior to and *necessary* for moral judgments. Further, I argue that this is an incoherent position because emotions do not provide the kind of specificity that the moral faculty relies on to make judgments of right and wrong. Given this negative conclusion, I turn in the third part to an alternative perspective that I believe can account for our judgments of right and wrong, while pointing to the significance of distinguishing between judgments and action. This alternative, based on

an analogy (made famous by John Rawls) with Chomskyan linguistics and the principles and parameters perspective most specifically, posits that we are endowed with a moral faculty that operates over the causal-intentional properties of an event, vis-à-vis the welfare of others. Given the many misunderstandings of this perspective, I first attempt to deflect criticisms that I do not believe are relevant. Next I state what I believe are the critical pieces of the linguistic analogy. And, last, I showcase how the analogy to language has opened the door to many new findings and predictions. I end the essay with a brief discussion of how our understanding of human nature, and especially the structure of our moral sense, may contribute in some small way to a life well lived.

2.1. A Bit of History

There is a long and rich tradition of discussion of our moral psychology, how it evolved and develops within each individual, how it differs between cultures, how we move from a description of what people naturally do in moral situations to what would be more desirable, and how legal and religious institutions might contribute to the creation of a more satisfactory moral universe. Philosophers have traditionally approached this problem using the power of logic and reasoning from examples, psychologists have used observational and experimental methods to chart how children acquire their moral sensitivities and ultimately put them in play in day-to-day life, and biologists have followed in the Darwinian tradition, contrasting the behavior of different species in morally relevant situations and attempting to infer what they think and feel from their behavior. These are all worthy traditions, and they have generated deep insights into what it means to be a moral creature. But nothing is ever peaceful in academia, and that is a good thing. The restless push us to think harder and, often, if we are lucky, in new and different ways. So, what's new?

The cognitive revolution forced the field to think in new ways about the mind. In particular, it forced an appreciation of controlled as opposed to automatic processes, and among the issues associated with automatic processes were questions concerning the principles guiding our intuitive reactions to the world and the extent to which the layperson had access to these principles. Thus, Chomsky pushed on the distinction between the kinds of grammatical rules that we learn in grade school—principles that are made explicit, learned in a controlled manner, often by association and rote memory—and the kinds of grammatical operations that are part of the child's native endowment, constrain the form of her expressed

language, and remain inaccessible to introspection except to the trained linguist. The same kind of distinction has been made by Kahneman in his work on decision making, including, as he did in his own Tanner Lectures, moral concerns.

On the moral front, however, the person most clearly responsible for the current zeitgeist is Jonathan Haidt, who picked up on Kahneman's general decision-making framework to make a pair of arguments. On the one hand, Haidt argued that much of what we perceive as rational, controlled decisions about moral rights and wrongs is illusory, mediated instead by a fast, automatic, and intuitive system. That is, we *think* we are deriving our dos and don'ts from principled reasons, but before we even turn on the reasoning engine, our intuitive system has fired off its decision. This, I believe, is the least controversial part of Haidt's argument. After all, there is no denying that our minds are equipped with the capacity to *both* reason from evidence and generate intuitive judgments driven by, well, intuitions of sorts. On the other hand, Haidt argued that our emotions provide the source of these intuitive judgments. This argument is, I believe, controversial, and I will address it head-on in section 2.2. Before doing this, let me offer an alternative that, in many ways, is most appropriately conceived as a necessary prerequisite.

Following on the heels of the Chomskyan revolution, Rawls drew an analogy between, loosely, our grammaticality and ethicality judgments. As Rawls reasoned, in the same way that we intuitively alight upon a judgment of grammaticality concerning a sentence in our native language, we generate a spontaneous, intuitive judgment of right and wrong for an event. Though Rawls clearly made explicit the distinction between this intuitive process and what he perceived as the kind of conscious deliberation that enters into reflective equilibrium, especially under the veil of ignorance, several commentators felt that he blurred or confused a more relevant distinction between descriptive and prescriptive concerns, a point that Mikhail has clarified in great detail. For now, I leave this historical-interpretive issue to one side, turning instead to more recent developments of the linguistic analogy, especially by Mikhail, Dwyer, and myself—the *moral grammarians*.

Of those who have picked up on Rawls's analogy, arguing in favor of the idea that all humans are endowed with a universal moral grammar that builds possible moral systems, the focus has been on clarifying the kinds of psychological distinctions that are in play, the transformational rules that might be necessary, and the extent to which the moral and linguistic

domains overlap. What is critical to appreciate at this point in the argument is that the sentimental and moral grammar perspectives are not alternatives in that both appeal to intuitive processes. Where they differ is in terms of the kinds of processes that motivate the intuition. As noted, Haidt and many others invoke emotional processes as central, whereas the moral grammarians push for causal-intentional processes. Needless to say, both processes could be going on, either in parallel or serially. For example, it is in principle possible that for every morally relevant event, we intuitively extract both the mental state and the emotional cause of action, evaluate the consequences of action for the welfare of others, and then derive a judgment of right or wrong. It is also possible that having made this judgment, additional emotional and causal-intentional analysis ensues, perhaps fueling revisions. To clarify why these distinctions matter, first in terms of descriptive ethics and second for prescriptive ethics, I turn next to an exegesis of the sentimentalist or Humean position.

2.2. O, Emotion, Where Art Thou?

Based on a wave of recent data, several authors have argued that emotions are necessary for moral judgment; some have even argued that they are both necessary *and* sufficient. A critical look at the evidence does not, I believe, support either conclusion. More specifically, based on both neurological and behavioral-psychological research, it is not possible to establish the *synchronic claim* that emotion partially or wholly constitutes our moral capacities, nor is it possible to establish the *diachronic claim* that emotion is necessary for the development of our moral capacities. This pair of negative claims does not constitute an all-out rejection of emotion in all aspects of moral thought and behavior. Rather, it is a specific claim about the timing and nature of emotional effects on moral judgment.

I first lay out some of the theoretically possible ways in which emotion might play a role in our moral psychology, then review the kind of evidence, both behavioral and neurobiological, that has been used to support a causal connection between emotion and moral judgment, and then end with new evidence that I believe puts this causal argument in jeopardy.

In brief, there are at least two ways in which emotion might play a role in moral thought and action. On the first view, dominant in today's discussions, emotions are triggered following the perception of an event, and then, consciously or unconsciously, modulate moral judgments. On the second view, something else guides the subject from the perception of an event to a moral judgment, but the judgment itself is either the linguistic

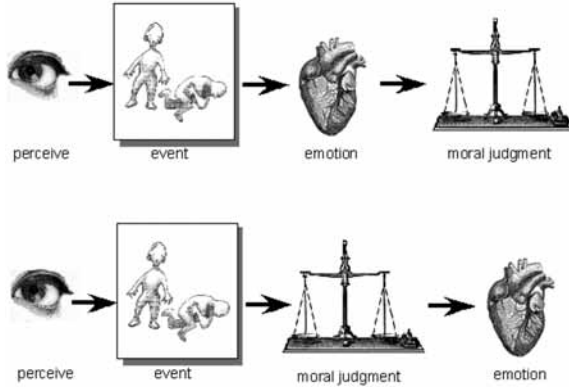


FIGURE 10. Two possible paths from the perception of an event to a moral judgment. In the top row, emotions are causally necessary and prior to moral judgments. In the bottom row, emotions are not causally necessary for moral judgments.

expression of the emotion or what follows from the emotional judgment (fig. 10). On the first view, it is of course possible for the moral judgment to again trigger an emotion, and for the emotion to feed back and modify the moral judgment. Thus, there could well be a cyclical choreography between emotion and moral judgment. Further, emotions triggered by moral judgment can then influence moral behavior or action. All of these possible models focus on issues of timing. They are silent on how emotions influence either judgment or behavior. What, therefore, is known about either the timing or the nature of emotional influences?

A brief encapsulation of the empirical landscape includes the following peaks of evidence: Violating moral norms triggers emotions, as evidenced by subjective impression and activation of neural circuits associated with emotion. Often, such emotions can inhibit morally reprehensible actions, as is the case when feelings of guilt and shame compel moral virtue. Some evidence for this modulating role of emotion comes from studies of psychopaths who have a greatly reduced capacity for empathy and guilt, and, apparently as a result, fail to inhibit their violent tendencies (but see below). Emotions can guide the process of moralization, as when disgust eliminates the human(e) element in the service of ethnic cleansing. And, last, morally relevant actions are often motivated by emotions, both early in human development and in phylogeny.

The problem with these observations is that they fail to distinguish the variety of ways in which emotions could be involved in moral thought and

action. There are four relevant points to make, each linked to behavioral and neurobiological evidence. First, while these studies reveal that emotions are *associated* with moral judgments, this does not mean that emotional responses *constitute* or are isomorphic with such moral judgments. To clarify the distinction between association and constitution, consider a parallel argument in a neighboring corner of the cognitive sciences, focused on the problem of “embodied cognition.” In its most extreme form, the embodied-cognition thesis holds that many concepts entail or are constituted by motor representations. The concept HAMMER is thought to consist of (1) features that typify hammers (for example, a grasping shaft, a hard end) *and* (2) the motor routines involved in goal-directed grasping and swinging. Support for this thesis often comes from neuroimaging data showing that, for example, the word “hammer” activates circuits classically associated with object categorization *and* circuits in the primary motor cortex. However, activation in the motor cortex does not license the conclusion that the concept HAMMER includes motor routines. Given the poor temporal resolution of neuroimaging, it is just as likely that the concept HAMMER activates circuits dedicated to object categorization that, in turn, activate circuits in the motor cortex. Distinguishing these hypotheses allows cognitive scientists to target motor routines to see whether they are *part of our concepts* or instead *stand in important causal relationships to them*. Analogously, it is essential that cognitive scientists target emotional mechanisms to see whether they *constitute* moral concepts or merely *stand in an important causal relationship* to them.

Second, while emotions can fuel the process of moralization, their central role may be restricted to an attention-grabbing function, one designed to draw the observer in and focus on the morally salient features of the environment, capturing attention and triggering distinctively moral cognition. Thus, for example, it is not that disgust tells us a moral wrong has transpired, but rather, that we should be on alert for an important social situation, one demanding our attention to work out why something happened and what the consequences might be. On this view, emotions alert the moral system, but do not provide an analysis that is sufficient for deciding moral rights and wrongs.

Third, most of the neurobiological data are correlational, and lack the relevant temporal information. More specifically, and as discussed more fully below, imaging studies can only show that areas associated with emotional processing are activated at some point in the evaluation of a moral dilemma, the temporal resolution is insufficient to show that they precede

and cause moral judgments. Similarly, studies of patients with adult-onset damage to emotion-related areas fail to distinguish between the thesis that emotions were necessary in forming moral judgments and emotions are necessary for online evaluation of moral dilemmas.

Last, and as I have argued before in *Moral Minds*, the fourth claim is most likely a correct characterization of the transformative properties of emotion with respect to our moral psychology. In particular, though I think everyone would agree that emotions can cause us to appreciate the ethical force of our judgments and motivate moral action—including the capacity to both empower and inhibit an act—the most prevalent views on emotions are that they are both necessary and sufficient for the possession of moral concepts; that making a moral judgment is nothing more, or less, than being in a particular emotional state; and that emotional circuitry is recruited in making some or even all moral judgments. In addition, a number of cognitive scientists have recently suggested that both empathy and sympathy emerge early in evolution and development, playing a pivotal role in kin-based altruism, reciprocity, and even pure altruism or what some describe as other-regarding preferences in the context of helping. As I discuss more explicitly below, the available behavioral and neurobiological evidence does not support this line of explanation.

To date, the primary evidence used to support the thesis that emotions represent the source of our intuitive moral judgments comes from studies manipulating, either explicitly or implicitly, people's current emotional state. For example, subjects provide more severe moral judgments when responding to moral dilemmas at a dirty desk, or when smelling a noxious odor. When subjects who are highly susceptible to hypnosis are hypnotically induced to experience disgust when confronted with a neutral word, they perceive moral transgressions as more severe in vignettes containing the hypnotically targeted word. Finally, subjects who watch a funny clip from *Saturday Night Live*, as opposed to a neutral control clip, not only report feeling in a more positive mood but generate more utilitarian responses to the trolley-footbridge dilemma (pushing the fat man off the bridge to save five others) but not to the bystander dilemma (turning the trolley onto one person to save five others). These are certainly interesting effects, but they do not show that emotions are constitutive of moral judgments.

Evidence from neurobiological approaches are also insufficient to show that emotions provide the essential source for our moral judgments. However, there are two sets of findings that make this part of the story both more complicated and more interesting.

Like the behavioral data, neuroimaging studies generate data that are both correlational and fuzzy with respect to the timing of different mechanisms. For example, judgments about morally significant claims (for example, “The elderly are useless”) show increased activity in the frontal polar cortex (FPC) and medial frontal gyrus, when compared to judgments about nonmoral claims (for example, “Telephones never ring”). Moreover, morally significant stimuli evoke increased functional connectivity between areas known to be involved in social decision making, reward evaluation, conflict resolution, and emotional experience (such as the left FPC, orbital frontal and anterior temporal and anterior cingulate cortices [ACC], and limbic structures such as the thalamus, midbrain, and basal forebrain). Last, moral dilemmas that push up close and personal contact (the fat man in the footbridge-trolley problem), in Greene’s terminology, recruit stronger activation from the emotionally relevant circuits than impersonal dilemmas (the bystander trolley problem) that can be computed by a colder calculus linked to outcomes.

On the basis of these and similar imaging results, Greene has proposed that moral judgment requires (1) a prepotent emotional response (subserved by circuits in the *medial frontal gyrus*, the *posterior cingulate gyrus*, and the *angular gyrus*) that drives moral disapproval, and reflective utilitarian reasoning (implemented in the *dorsolateral prefrontal cortex* [DLPFC]). Although these systems often produce convergent outputs, outputs diverge in personal-moral dilemmas, generating conflict (evidenced by increased activity in the ACC) that must be resolved by higher-cognitive control circuits in the anterior DLPFC. Converging data from neuroeconomics suggest that in bargaining games involving the assessment of fairness, emotional circuits associated with the insular cortex are significantly activated in the face of perceived inequities. Several authors have argued that increased activity in this area is indicative of the sensitivity to norm-violations implicated in nonconsequentialist, deontological judgments.

As Mikhail has noted, it is hard to believe that anyone “would doubt or deny” the conclusion that “some perceived deontological violations are associated with strong emotional responses.” What we are after, but generally lack, are studies that show a causal link between emotion and moral judgments, accompanied by the timing and locus of such effects. Stated differently, the fact that neural circuits classically associated with emotion activate when people process moral scenarios does not enable us to favor the interpretation that (1) emotions are integral to *moral computation* over the interpretation that (2) emotions result from these computations.

Neuropsychological studies of patient populations hold out the hope of stronger evidence with respect to causality, although even here we must be cautious. The fact that an area X , when damaged, is associated with deficit Y , allows only a weak causal claim. Specifically, from these kinds of data we are licensed to conclude that area X is involved, in some way, with generating normal processing for capacity Y . Its involvement is uncertain, however, as X could, for example, modulate upstream a critical piece of the circuitry, or act as a gain function on a computation that has already arisen. For example, perhaps the computation necessary to generate Y has occurred before it gets to area X , but when X is damaged, the output is so weak that we observe a deficit. With these points in mind, I turn to two studies of patient populations, the first consisting of patients with clear, anatomically localized deficits to the ventromedial prefrontal cortex (VMPC) and concomitant behavioral deficits in decision making, the second with unclear anatomical damage but clear behavioral deficits—psychopaths.

Patients with adult-onset, bilateral damage to the VMPC exhibit four notable characteristics: (1) based on skin-conductance measures, they show flattened social emotions such as empathy and embarrassment; (2) they are unable to redeploy emotional representations previously associated with punishment and reward; (3) when they engage in actions that yield beneficial or costly financial returns (for example, the Iowa gambling task, the ultimatum game), they show deficits in their ability to anticipate future outcomes, punishments, and rewards; and (4) they are impulsive, as measured by standard executive tasks that tap into inhibitory control. The general conclusion that many derive from this work is that emotional processes play an integral role in decision making and, more specifically, that emotional processes implemented in the VMPC are critical to moral decision making.

Further evidence for the supporting role of emotion in moral decision making comes from patients with frontotemporal dementia, a disorder that is caused by the deterioration of prefrontal and anterior temporal cortex. These patients show blunted emotional profiles, disregard for others, and a willingness to engage in moral transgressions.

With these patterns observed, Koenigs, Young, and colleagues set out to look more closely at the nature of the deficit observed in VMPC patients. Two issues in particular motivated this study. First, though VMPC patients clearly show diminished social affect, and clearly act in a way that suggests a diminished understanding of the sociomoral domain, the ma-

majority of work on these patients focused on their actions as opposed to their judgments. Second, given the vastness of the moral domain in terms of the psychological processes it engages (for example, issues of utilitarian versus deontological concerns, virtues, harming versus helping, and so forth), it was not entirely clear how VMPC damage alters, if at all, these different processes. To address these issues, we presented a well-studied population of VMPC patients with a variety of social vignettes, some moral, some not. Among the moral cases were stories that, in previous work by Greene, had been labeled as *personal* or *impersonal* using the description above. Critically, although all of these moral dilemmas were judged as emotionally salient, all of the personal cases were judged to be more emotionally intense than the impersonal ones.

When Koenigs and colleagues analyzed the pattern of judgments, and in particular the extent to which subjects stated that it was permissible to act (such as flipping the switch in the bystander case, pushing the man in the footbridge case), the VMPC patients were indistinguishable from healthy controls as well as brain-damaged controls (that is, individuals with damage to brain regions outside of those believed to be relevant to our moral psychology) for nonmoral cases as well as impersonal moral dilemmas (fig. 11). VMPC patients did, however, generate significantly higher permissibility judgments for the personal moral dilemmas. But even here, the class of personal dilemmas presented a highly heterogeneous set of responses for all three groups. Thus, based on an insight by Young, a further cut was made in the data set, mapping roughly onto actions that were self-as opposed to other-serving. For most of the self-serving cases, for which there were no differences between groups, subjects answered quickly and agreed that the act was forbidden. For example, all subjects agreed that it was not permissible for a teenage girl to kill her newborn infant. For most of the other-serving cases, VMPC patients were much more likely to support the act, thereby favoring the utilitarian outcome. For example, they were more likely than the controls to say that it was permissible to push the fat man in the footbridge case.

Overall, then, these data show that even though VMPC patients experience a flattened socioemotional profile, they nonetheless judge most moral dilemmas (that is, impersonal moral dilemmas and personal dilemmas where *harm to another* is pitted against a *benefit to self*) as do healthy individuals. Moreover, given the limited range of cases on which VMPC patients deviate, it is plausible that they fail to treat the morally salient features of high-conflict dilemmas as *morally salient*. If this interpretation

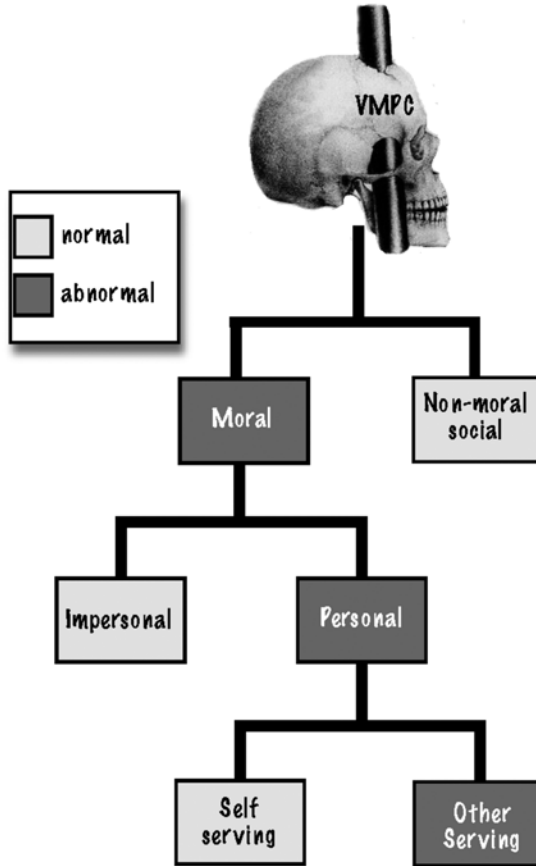


FIGURE 11. The moral landscape and the pattern of normal (light gray) and abnormal (dark gray) responses by patients with damage to the ventromedial prefrontal cortex.

is correct, then moral cognition would yield deviant outputs as a result of deviant inputs, rather than as a result of a deficit in moral processing per se; this parallels the interpretation we offered above for the hypnotism study. In this case, the process of moral evaluation would remain intact; however, the flattening of negative emotion would yield more permissible moral judgments because of a failure to focus on the *antecedently* morally salient features of the scenarios.

We can add to this general account—one in which emotions emerge out of moral judgments rather than providing their source—by drawing on recent work with psychopaths. Like VMPC patients, psychopaths also present with flattened social emotions, especially empathy, guilt, and

remorse. And like VMPC patients, there also appear to be significant problems with inhibitory control, a deficit that manifests itself in social (for example, violence) and nonsocial contexts (such as reversal learning tasks). Unlike VMPC patients, however, psychopaths present with high levels of violent volatility as well as the desire for extreme manipulation.

A crucial finding in the literature on psychopathy is James Blair's careful studies showing that psychopaths fail to perceive a difference between moral and conventional violations. This distinction, first articulated by Turiel in the 1980s, and replicated by many others, especially Judy Smetana, is based on the idea that in the social domain, there are different kinds of conventions, only some of which are moral. Though there is controversy concerning the borders of this distinction, there are several generic properties that distinguish conventional from moral transgressions. Thus, in contrast to conventional transgressions, moral transgressions are more severe, cross-culturally universal, closed off from the arguments of authority figures, and emotionally intense. For example, if a teacher tells her students that, unlike her present class, students in other countries never raise their hands when they ask questions, and that, from now on, they will ask their questions without raising their hands as well, most students find this okay, don't get upset, and go with the flow of the teacher's desires. Take the same general setup, but now the teacher says that instead of talking with their classmates about a problem, they should just turn around and smack them; the tables turn: children in other countries are barbaric, and the teacher has gone mad. When cases like these are presented to psychopaths, they fail to see a difference as evidenced by their tendency to judge both conventional and moral transgressions as wrong.

The psychopath's responses on the moral-conventional distinction are, however, a bit more complicated and interesting with respect to the source of the deficit. In particular, though both adult incarcerated psychopaths and juveniles diagnosed with psychopathy tend to allow more moral transgressions than healthy normals, adults are more permissive when there are no clear rules, whereas juveniles are more permissive irrespective of explicit rules. Thus, something clearly changes over development, and at this point, it is not clear whether the change is due to maturation of prefrontal cortices (critical for linking social decision making with emotion), experience with rules handed down by the local culture, emotional maturity, or some combination of these factors and others.

As robust as these results are, they leave open a question about the psychopath's moralspace: is their apparent inability to see a distinction

between conventional and moral transgressions due to a deficit in their moral knowledge or to an understanding of what is expected in such cases? More specifically, what do psychopaths do when they confront a moral dilemma in which there is no clear answer, no clear obligation to pick one route or the other in terms of moral permissibility? To address this question, I teamed up with two forensic psychologists, Maaïke Cima and Franca Tonnaer, presenting psychopaths with a series of moral dilemmas.¹² To make the contrast with VMPC patients as robust as possible, we used the same personal moral dilemmas, as these were the cases that showed the most striking differences from healthy controls. In addition, we picked two control groups that we believed would enable the clearest interpretation. Specifically, since psychopaths are largely male, we selected two all-male controls, both matched for age with the psychopaths. One group was composed of healthy males with no known clinical deficits, and the second group was composed of nonpsychopathic delinquents. Psychopathy was assessed using a standard battery of questions (that is, the PCL-R test). Each group was given the same impersonal and personal moral dilemmas presented to the VMPC patients. Though all groups showed variation in their responses to these dilemmas, with consistently fewer affirmative responses to the question “Would you *X*?” for personal than impersonal moral dilemmas, there were no significant group differences for either the personal or the impersonal dilemmas. Moreover, contrasting self- and other-serving personal moral dilemmas also failed to reveal a group difference (fig. 12).

In sum, though psychopaths clearly show deficits in social emotions and inhibitory control, as well as in their capacity to distinguish moral from conventional violations, they do not show differences in their judgments of either impersonal or personal moral dilemmas. Given the highly varied and complicated structure of the dilemmas presented, and the equally complicated pattern of neural activity elicited by such dilemmas in Greene and colleagues’ work, it appears that the psychopathic mind functions normally. Put differently, whatever knowledge is necessary to navigate within the morphospace defined by the impersonal and personal dilemmas presented, psychopaths present sufficient competence. Moreover, whatever role emotion plays in our moral psychology, it cannot be necessary for these moral judgments. Not only do psychopaths make judgments that are indistinguishable from healthy normals, but their

12. M. Cima, F. Tonnaer, and M. Hauser, “Psychopaths Know Right from Wrong but Don’t Care,” *SCAN* 5 (2010): 59–67.

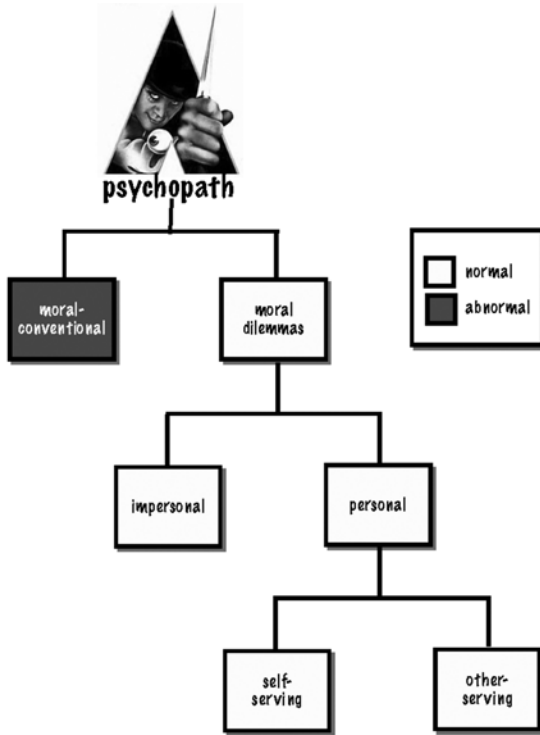


FIGURE 12. The moral profile of a psychopath. White boxes present responses by psychopaths that appear normal relative to various control populations, including healthy age-sex matched subjects, as well as nonpsychopathic delinquents. Black boxes present abnormal responses.

judgments are also indistinguishable from nonpsychopathic delinquents who are, presumably, emotionally damaged at some level. These data suggest that the most significant role for emotion is in motivating morally relevant action, which may or may not coincide with moral judgments. In the case of the psychopath, and, presumably, the nonpsychopathic delinquent as well, abnormalities in emotional processing lead to immoral behavior; they may also lead to a lack of appreciation or concern for their judgments. And from their immoral behavior there is no subsequent emotional response, either in terms of negative emotions such as guilt and remorse, which serve to deter repeat performances in healthy individuals, or in terms of positive emotions such as empathy, which serve to motivate virtuous behavior.

Much, of course, remains unclear. For example, though psychopaths show deficits in processing or experiencing sociomoral emotions, they appear to show intact processing of other emotions such as anger and disgust. Since we have a rather poor understanding of how specific emotions link up with understanding of specific moral scenarios, it is not possible to assess whether the *preserved* emotional experiences of a psychopath are sufficient to guide moral judgments. Further, although psychopaths present relatively normal moral judgments when presented with moral dilemmas, they may show little to no appreciation of why these judgments matter, or why anyone should care. Thus, a lack of emotion may cause deficits in both moral appreciation and moral action. These are questions for future work.

Before leaving the sentimentalist position, let me end with a comment about the *diachronic* issue: that is, are the emotions necessary for the acquisition of moral concepts and the development of moral judgment? In some of the early and elegant work on moral development, Hoffman appealed to the affectively laden tools used by parents to convey social rules and correct behavior in arguing that emotions are developmentally necessary for moral judgment. It is not clear, however, whether the role of corrective measures speaks in favor of the necessity of emotions. First, the nature of moral correction is usually so vague as to be incoherent with respect to the moral architecture that is both necessary for trafficking in the moral domain and that the child ultimately acquires. Second, although there is extremely little known about the kind of input the child receives during development, and how she distinguishes morally relevant experiences from irrelevant, it would appear that most rule-based correction is directed toward conventional transgressions (“Take your finger out of your nose!”) as opposed to moral ones (“Don’t kill your brother.”). Further, it seems that parents most often accompany their *don’t*s with an obligatory emphasis of *must* in conventional as opposed to moral cases, such as “You must clean your room, eat your broccoli, stop picking your nose” but not “You must stop hitting your brother, lying, breaking promises.” These are only speculations. As in studies of child language acquisition, it is clear that we need serious studies of the actual input the child receives in the moral domain, the extent to which negative evidence is present, the impact of correction on subsequent understanding, and so forth. Until we obtain such data, we will be on uncertain ground with respect to explaining what does or does not affect the child’s moral development, at least in terms of competence.

An alternative ontogenetic hypothesis relies on the moral deficiencies of psychopaths. Blair claims that in normally developing children, emo-

tional circuits facilitate negative reinforcement for actions that generate distress cues. Psychopaths lack these emotional circuits, and Blair argues from this fact to the claim that emotion is the developmental source of our moral concepts and that psychopathy is a developmental consequence of an early emotional deficit.

While psychopaths fail to distinguish moral from conventional transgressions, treating conventional violations as less permissible, more serious, and authority independent (that is, as moral transgressions), such data do not speak to either the source or the content of a psychopath's moral cognition. This pattern of response is equally well accounted for by a cold, calculated rationality, designed to get "off the hook" and to say what others want to hear. Moreover, as Blair himself has shown using age-matched psychopathic and nonpsychopathic juvenile delinquents, even psychopathic juveniles draw the moral-conventional distinction (though less pronounced in psychopathic juveniles) and make just as many references to welfare considerations as do nonpsychopathic controls (though psychopathic juveniles were less likely to ascribe moral emotions to others). These data suggest a developmental trajectory for psychopathy, but contrary to what one would predict if emotion is developmentally necessary for acquiring moral concepts. Psychopathic juveniles apparently lose the capacity to distinguish moral from conventional violations over the course of development. So, perhaps the deficiencies in the moral psychology of the psychopath are a developmental consequence of antisocial behavior, instead of the other way around. As Adrian Raine argues, a life filled with antisocial behaviors may modify an individual's moral psychology, allowing for the justification of immoral behaviors and reducing cognitive dissonance. But if this is true, the moral cognition of psychopaths is deviant as a result of deviant inputs rather than as a result of a deficiency in moral processing.

A final piece of evidence against the diachronic claim comes from an exploration of the moral psychology of early-onset VMPC patients. In contrast with the adult-onset cases discussed above, individuals with early-onset lesions to VMPC are unresponsive to punishment, lack inhibitory control, generate behavioral deficiencies in moral and prudential domains, and show emotional deficits in guilt, remorse, and empathy. Such patients also fail to acquire moral concepts, justifying their behavior by appeal to the egocentric desire to avoid punishment. However, such data tell us only that the acquisition of moral concepts is downstream from some social-emotional mechanisms. In order to establish that emotional processes are constitutive of moral cognition, we would need a much clearer picture of

the precise deficits present in early-onset VMPC patients, testing them on the same battery of dilemmas used in the adult study. Given the rarity of this disorder, it is unclear whether the absence of moral competence is a deficit in the acquisition of social rules produced by a lack of positive feedback or even a result of deviant behavior that inhibits the maturation of moral cognition.

Admittedly, this section has been one long critique, targeting some of the fundamental problems with the sentimentalist position of emotions-as-source. What is needed, and what I owe, is an account that can explain the relationship between people's judgments and the complicated dilemmas presented. I turn to this next.

2.3. On Moral Grammars, Universality, and Cultural Variation

The famine crisis in Bangladesh covered the airwaves and papers in the 1960s and early 1970s. To keep the Bengali refugees alive for one year would have cost the world approximately \$750 million. Britain was one of the major contributors, giving about \$25 million in one year. This may seem like a healthy contribution for one nation, but in that same year, Britain contributed close to \$500 million to help the French build the Concorde jet. As philosopher Peter Singer remarked, "The British government values a supersonic transport more than thirty times as highly as it values the lives of nine million refugees."

The situation today is hardly better. The World Bank estimated that out of approximately 6 billion people on earth in the year 2000, almost half fell below the poverty line. Poverty translates not only to hunger, but illnesses and insufficient medical aid. In 2005, the United Nation's World Food Programme projected that it would cost just over \$3 billion to feed 73 million hungry people, leaving an additional 800 million people in a state of starvation. Providing relief for the remaining numbers, and ending world hunger for 2005, would run the globe an additional \$35 billion, bringing the tab up to about \$40 billion. The United States spends about \$40 billion each year on gambling, a superfluous activity that no one needs. If everyone stayed away from the slot machines and poker tables for just one year, voilà, hunger relief for all. Alternatively, and as Singer has argued, carving off a mere 1 percent of earnings from the world's richest individuals would do the trick.

When statistics such as these are trotted out without labeling the countries—for example, in response to country A's need for \$500 million in relief funds, country B gave \$10 million but also spent \$80 million on

the production of several blockbuster movies—it is hard to imagine anyone having the intuition that such policies of resource distribution are permissible. In fact, the policies seem wrong—morally wrong, that is. Why, then, has the situation remained unchanged? Why are we so incapable of doing the morally right thing? Why do our intuitions fire one way and our actions another? When are we morally obligated to help? Can we develop policies that harness the right intuitions or correct for unfortunate ones given the current climate?

Singer provides a simple principle to guide the psychology of obligatory aid: *If it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it.* What is crucial about this principle is that it links the psychology of moral obligation to the cold calculus of cost-benefit analysis and the systems of motivation. And it creates this link without making reference to a particular group of people and their relationship to the morally responsible agent. Nothing in the Singer Principle depends upon whether the individual or individuals in need are neighbors or foreigners, near or far. Further, nothing hangs on whether there are one or more potential contributing agents to the cause. Everything hangs, however, on the phrase “without sacrificing anything of comparable moral importance.” This is a sticky phrase, one that has constantly confronted utilitarians such as Singer. The stickiness is due to one word—“comparable”—and its frame of reference: *comparable* with respect to what standard and whose standards?

To highlight the challenge of making Singer’s principle do some real work, consider three cases. Are we morally obligated to give money to an aid organization to purchase rehydration salts for children in sub-Saharan Africa instead of buying a candy bar? Saving one or more children from death due to dehydration is unquestionably worth the personal sacrifice of junk sugar. Our moral calculus should compel our motivational systems to act and do the morally right and virtuous thing: give to the aid organization. Here we push against the problem of bypassing the short-term but small benefit for the delayed but large benefit for both self and other.

What about our moral obligation to fund this relief program instead of sending our children to college? Our children do not *need* an education for survival. In the elite American universities, tuition in the year 2005 exceeded \$200,000 for one student, a sum that would do wonders for relief programs. But knowledge is a great thing to own, and colleges make this possible. Everyone, universally, agrees that education is important. But few, if any, think that an education is more important than surviving.

Finally, consider the trolley problems that put into play the moral importance of an individual's life. The bystander can prevent five people from dying by flipping a switch or pushing a heavy man onto the tracks. Most people think it is permissible to flip the switch but forbidden to push the man. In seeing the switch case as permissible, we see one person's life as less important than five. Using the Singer Principle, it is in the agent's power to prevent something bad from happening (killing five), without thereby sacrificing anything of comparable moral importance (one life is not as important as five lives); therefore, the agent ought, morally, to flip the switch. In the pushing case, we flip the argument around, essentially saying that one person's life is of equal (or greater) moral importance compared to five people's lives, and, thus, we should not push the man. Filling in the principle, it is in the agent's power to prevent something bad from happening (killing five), but killing one entails sacrificing something of comparable moral importance and, thus, the agent ought not, morally, to push the man. As stated, the Singer Principle cannot arbitrate between these cases. How do we decide what counts as comparable in "moral importance"? Are we missing the essential parameters that modulate the outcome of our decision? Should we allow for a liberal plurality of views under some circumstances? No easy answers here.

This brief discussion of world hunger takes us back to the competence-performance distinction that Chomsky made famous in linguistics and that Rawls's linguistic analogy invites: how do our intuitive judgments of permissible, obligatory, and forbidden actions clash with what we would in fact do? The Singer Principle is an idealization of what we ought to do. It is a beautifully simple and clear prescriptive principle. It is, of course, possible that at some level of abstraction it is also a descriptive principle that is part of our moral grammar. Everyone, it seems, has the intuition that cost-free rescue is morally obligatory: we must save the drowning baby from the bathtub even if we get wet, and we must give our candy bar to a starving child even if we were looking forward to a delicious snack. These situations are easy, and the idealized principle works beautifully. But the world is ugly.

From the triggering of a distinction such as the Singer Principle to the implementation of an action, there are many mind internal and external processes that intervene and contribute to our behavior including the motivational and cost-benefit systems. For example, the Singer Principle states that we should be as willing to provide aid to our neighborhood's man on the street whom we see each day on our way home from work as to the starving man in Somalia, presented on the cover of an Oxfam card.

One ugly fact, however, is that the human brain is equipped with a powerful, rapid, unconsciously operative, prejudicial system that privileges the in-group against the out-group. In fact, our brains are more likely to build an association of a fearful stimulus with a member of a different race than with a member of the same race. Paralleling results with our primate cousins, we have evolved a brain that is prepared to perceive other groups as negative and fear inciting. The neighborhood's man on the street is in, whereas the man in Somalia is out. And that's just for starters. The man on the street is out relative to our friends. The man on the street is also out in terms of our selfish interests in owning more superfluous stuff.

Our prescriptive ethics tells us that all of these factors represent noise in the system and, thus, should not count. Human nature says that they do count, and have for millions and millions of years before our emergence on earth. As philosopher Richard Rorty put it:

To get whites to be nicer to blacks, males to females, Serbs to Muslims, or straights to gays... it is of no use whatever to say, with Kant: notice what you have in common, your humanity, is more important than these trivial differences. For the people we are trying to convince will rejoin that they notice nothing of the sort. Such people are morally offended by the suggestion that they should treat someone who is not kin as if he were a brother, or a nigger as if he were a white, or a queer as if he were normal, or an infidel as if she were a believer.

We all live with the tension between in-groups and out-groups. Whether or not we engage with this tension is another matter.

These are hard problems. They have been debated for decades by philosophers, politicians, lawyers, lobbyists, and presumably countless families eating dinner in the developed world. They raise fundamental challenges for each of us as we contemplate the ingredients that enter into a moral life. In this last section, I want to review the main arguments for thinking about morality as Chomsky and other generative linguists have thought about language. I then push as hard as possible on this theory, showing where I believe it has succeeded and failed, and what lies on the horizon, both near and far. Finally, I hail a few *mea culpa*s for all of the fascinating and complicated issues that I have omitted from discussion, and defend why this restricted coverage was necessary. It is a strategic apology based on a view that is commonplace in the natural sciences: depth of understanding comes only from idealization away from the complexities of the world writ large, focusing more narrowly on a corner that can hopefully be explained in a rich way.

In 1998, Rawls published his final book, titled *Justice as Fairness*. It was an update of *A Theory of Justice*, covering the wide range of critiques that appeared over a thirty-year period. Most of the core ideas are intact, including the anchoring concept of justice as fairness. The linguistic analogy is, however, gone. Although there is an interesting history here, elegantly reviewed by Mikhail, I will skip it in order to breathe new life into the analogy between language and morality, vindicating Rawls's intuition, highlighting several exciting empirical observations that have emerged since both my own writings, and those of my fellow moral grammarians. Consider the following a sequel: Rawls reloaded!

In *Justice as Fairness*, Rawls engages in a style of argumentation that dates back to Galileo and Newton, has been preserved to this date by many physicists and chemists, and, most relevantly, has been championed by Chomsky, inspired to some extent by the molecular studies discussed in Lecture I.¹³ For Galileo, the nature of nature is perfect. Theories that tackle aspects of nature's perfection by abstracting away from perceived irrelevancies are preferred over those that try to accommodate everything. It is a perspective in which theory, and especially idealization away from the complexities of the world, often trumps objectively acquired data. Albert Einstein, who was renowned for his lack of interest in confirming or disconfirming evidence, once commented that if Sir Arthur Eddington had not found supporting observations for the general theory of relativity, "I would have been sorry for the dear Lord—the theory is correct." To understand the world, we need to idealize away from it, focusing on a small corner, allowing the beauty of our ideas to surface even if they fail to explain certain pieces of data. For Chomsky, the computations running language may represent an optimal and beautiful solution to the constraints imposed by our thoughts and the machinery we use to express ourselves. How this system is put to use in the world is ugly, a complicated reflection of mind internal and external factors. Understanding comes from idealization, which includes stepping away from commonsense descriptions, as well as the richness and complexity of what we see. Rawls advanced the same Galilean stance of abstraction and idealization:

In using the conception of citizens as free and equal persons we abstract from various features of the social world and idealize in certain

13. The Galilean style and the beauty of theory (Noam Chomsky, "Minimalist Inquiries: The Framework," in *Step by Step*, edited by R. Martin, D. Michaels, and J. Uriagereka [Cambridge: MIT Press, 2000], 89–155; Steven Weinberg, "The Forces of Nature," *Bulletin of the American Academy of Arts and Sciences* 29, no. 4 [1976]: 13–29).

ways. This brings out one role of abstract conceptions: they are used to gain a clear and uncluttered view of a question seen as fundamental by focusing on the more significant elements that we think are most relevant in determining its most appropriate answer. Unless explicitly stated otherwise, we do not try to answer any question except the fundamental question (of political philosophy)... what is the most acceptable political conception of justice for specifying the fair terms of cooperation between citizens regarded as free and equal and as both reasonable and rational, and (we add) as normal and fully cooperating members of society over a complete life, from one generation to the next?¹⁴

I have taken a similarly Galilean route in trying to explore the linguistic analogy. It both recognizes and appreciates the vast complexity of the moral domain while realizing that any attempt to explain all of it is hopeless. Instead, it takes the more modest approach of trying to explain one small corner of our moral psychology. In so doing, I have made a number of simplifying assumptions. The hope is that by using a microscopic lens, myopically focused on a small corner of the phenomenon, that we have gained some explanatory power and depth. I believe we have.

At the heart of the linguistic analogy is a core set of ideas that must be addressed, independently of whether morality as a domain of knowledge is anything like language. To understand the moral domain, we must provide a rich description of the principles that underlie the mature state of knowledge, explain how this knowledge is acquired in individual development, and dissect the anatomy of the system in order to define which components are uniquely human and uniquely moral. Though most of the evidence that I presented in *Moral Minds* was collected *without* this motivating framework, it has since generated several new findings that would not have been collected otherwise.

Across a number of different social situations in which moral concerns surface—ranging from harming to helping others—humans deliver judgments based on intuition as opposed to principled reasoning.¹⁵ In cases

14. John Rawls, *Justice as Fairness* (Cambridge: Harvard University Press, 1998), 5.

15. For clarity, it is important to distinguish between our judgments, our appreciation of such judgments, and the actions that ensue. It is commonly claimed that *real morality* entails only action, what we do in the course of engaging with ethics. Though action is undoubtedly critical, we must not underestimate the frequency with which we make pure judgments without action: reading novels and the newspaper, watching the news, hearing accounts from friends, and so forth.

where cross-cultural evidence is available, we observe both universality as well as constrained variation. Across cultures, straightforward deontological or utilitarian concerns fail to account for people's judgments. No culture strictly holds to the rule that *killing is wrong*. And the lack of adherence to this rule is not simply because we all go postal from time to time, enraged by envy or a lover's infidelities. All cultures hold to a principle that dictates, under parametric variation, when killing is permissible, obligatory, or forbidden. Let me clarify by describing recent work that I carried out in collaboration with Linda Abarbanell on a rural Mayan population. In order to contrast data from a small-scale society with data collected from the Moral Sense Test (MST) on the Internet,¹⁶ it was necessary to translate the dilemmas into Tselal and then back to English. It was also necessary to modify the forbidden → obligatory response scale to one ranging from very bad → very good; this was necessary because Tselal lacks words for *forbidden* and *permissible*.

We presented Mayan adults with several dilemmas, aimed at testing the hypothesis that, like our Western Internet sample, they perceive means-based harms as worse than side effects, and actions as worse than omissions. In the first round of dilemmas, Mayan subjects judged means-based harms as worse than side effects, but they judged action-based harms as harshly as omission-based harms. Though we were pleased by the replication of the means–side effect distinction, we were surprised by the failure with actions and omissions. Although there certainly are studies showing that under certain circumstances, subjects perceive no difference between actions and omissions (for example, when the individuals in the scenario are highly familiar), the general finding in the literature, and confirmed with large sample sizes by our own work, is that folk intuition puts actions as worse than omissions. We thus presented a new population of Mayan subjects with additional action-omission dilemmas. Some of these dilemmas had been presented to less traditional Mayans living in a city and with more formal education, whereas other dilemmas had been presented in the same format and with the same scale on the MST, or to young English-speaking children; in all of these cases, subjects consistently judged actions as worse than omissions. This new population of Mayans, however, failed to perceive a difference. To probe this failure further, we presented both a new set of means–side effect cases as well as a pair of action-omission dilemmas in which the question shifted from moral permissibility to causal responsibility. That is, based on previous work showing that causal trans-

16. See <http://www.moral.wjh.harvard.edu>.

parency is largely responsible for driving the action-omission distinction (that is, the agent acting to bring about harm is seen as more causally responsible for the outcome than the agent omitting this action), we wanted to see if the failure with actions and omissions was mediated by a failure to perceive causal responsibility. Mayan subjects readily perceived means-based harms as worse than side effects, and also perceived agents involved in action-based harms as more causally responsible than agents involved in omission-based harms.

What these results suggest is that components of our moral computation are universal (for example, the means–side effect distinction), running through different cultures, whereas others (such as action-omission) are open to variation; with respect to the linguistic analogy, this may map to universal *principles* and optional *parameters*. In particular, the Mayan data raise two interesting possibilities with respect to the source of variation. On the one hand, perhaps something particular about Mayan society forces omissions to be treated seriously, in much the same way that Good Samaritan laws are explicit in countries like France. Although we could not find anything in Mayan ethnographies to suggest an explicit social norm such as the Good Samaritan law, it is, of course, possible that such norms exist. An alternative possibility is that the failure to perceive a meaningful difference between actions and omissions is due to general properties of small-scale societies as opposed to Mayan culture specifically. This view is supported by Haidt and Baron's work on how social roles, and especially familiarity, can eliminate the perceived difference between actions and omissions.¹⁷ Thus, the fact that educated Mayans living in a big city perceive the action-omission distinction could be due to their education, the anonymity of a big city, or some combination. What is needed, following the lead of cross-cultural studies of language, is to pick apart which aspects of the input are most relevant to the observed differences, to establish when in development such input affects the child's moral psychology, and to assess the plasticity of the developmental program for acquiring a moral system. As noted within the principles and parameters perspective, this view makes sense for morality only if there is variation but it is constrained.

The success of a theory to explain existing phenomena is only the first step. Generating novel predictions is the next or parallel step. Riding on the richness of work in linguistics, the theory discussed here generates a

17. J. Haidt and J. Baron, "Social Roles and the Moral Judgment of Actions and Omissions," *European Journal of Social Psychology* 26 (1996): 201–18.

wide range of questions and predictions concerning the operation of the moral faculty. For some, this may seem depressing. Personally, it is the ultimate rush. As long as one sees that the myriad questions and predictions are genuinely interesting, and worthy of study, the potential for deep understanding is extraordinary. Let me note, however, that by making an analogy to language, I am not buying into an assumption that language and morality work in precisely the same way, based on the same core set of processes and representations. This would be absurd. The basis of the analogy is first to raise questions about the possibility of descriptive principles underlying the mature state of moral knowledge and, if operative, to characterize the acquisition of such principles. We have virtually no understanding of these issues. Gaining an understanding will help us reveal the ways in which morality and language rely on comparable computational processes, and the ways in which they differ.

Consider three examples of relatively unstudied problems in the moral domain, each opened up by the analogy to language. First, few within the field of moral psychology have seriously entertained the possibility that the principles currently used to describe our moral judgments are less like the rules we learn in grammar school and more like the rules or principles that linguists bring to bear on the problem. The former look only to the surface of the problem; the latter dig deeper, assuming that there are abstract and operative principles lurking. I return to this issue below.

Second, I know of no study focused on the question of critical periods in moral development and, especially, whether the acquisition of the first or native moral system is qualitatively different from the acquisition of a second and later system. This is surprising in part because it is such an obvious question once one draws the analogy to language. But it is also surprising because anyone who has traveled outside of their local turf has experienced the challenges of working out another culture's social norms. And there are natural experiments as well: every time a child, from infancy through to the teenage years, is adopted and moved to a different culture, we have the perfect pre- and postexperimental design. In studies of language, this approach has led to many important discoveries, most recently by Jesse Snedeker, who has found that adopted children go through the same developmental milestones in their new language as do children acquiring a native language for the first time; the main difference is that the adopted children go through each milestone much more quickly than do nonadopted children.

Third, all of the work on the evolution of morality in animals has focused on their behavior, on what individuals do when they have the opportunity to help or harm another. There is no work asking animals to judge whether particular actions are permitted or forbidden. Although this is a nontrivial methodological problem, one could imagine an experiment in which an individual watches a series of actions that are “allowed,” meaning that they are consistent with species-typical social practices such as: high-ranking animals taking away food from low-ranking animals, adult males mating genetically unrelated adult females, and individuals sharing an abundance of food by giving food calls. After repeatedly playing these film clips, present new clips that reveal novel but allowable interactions versus novel but forbidden interactions; the latter might include low-ranking individuals taking food away from high-ranking individuals without a fuss and unforced mating among close genetic kin. Longer looks to the transgressions would not give us evidence of permissibility judgments. They would, however, get us closer to their expectations, and, thus, to their competence in judging the outcomes of morally relevant interactions. Even if this design falls short, the main point holds: students of animal behavior with an interest in the evolution of morality must think about the distinction between competence and performance, recognizing that what an animal does may not map onto their perception of the same situation prior to action.

To evaluate where things stand, I reproduce below a list of features that I proposed in *Moral Minds* to capture the anatomy of the moral faculty. As noted at the time, these are features that would be expected if the linguistic analogy is taken in its strongest form, with morality and language being isomorphic in important respects. What the theory allows, but critics have often ignored, is a weaker position, one that takes the analogy as a heuristic for generating novel questions and leaves open the distinct and unsurprising possibility that the two domains differ in many important ways:

1. The moral faculty consists of a set of principles that guide our moral judgments but do not strictly determine how we act. The principles constitute the universal moral grammar, a signature of the species.
2. Each principle or set of principles generates an automatic, rapid, and confident judgment concerning whether an act or event is morally permissible, obligatory, or forbidden.
3. The principles are inaccessible to conscious awareness.

4. The principles operate on experiences that are independent of their sensory origins, including imagined and perceived visual scenes, auditory events, and all forms of language—spoken, signed, and written.
5. The principles of the universal moral grammar are innate.
6. Acquiring the native moral system is fast and effortless, requiring little to no instruction. Experience with the native morality sets a series of parameters, giving birth to a specific moral system.
7. The moral faculty constrains the range of both possible and stable ethical systems.
8. Only the principles of our universal moral grammar are uniquely human and unique to the moral faculty.
9. To function properly, the moral faculty must interface with other capacities of the mind (for example, language, vision, memory, attention, beliefs), some unique to humans and some shared with other species.
10. Because the moral faculty relies on specialized brain systems, damage to these systems can lead to selective deficits in moral judgments. Damage to areas involved in supporting the moral faculty (such as emotions, memory) can lead to deficits in moral action—of what individuals actually do as opposed to what they think someone else should or would do.

The first four features concern the mature state of knowledge, the next three concern acquisition, and the final three concern uniqueness, both over evolution and across domains of knowledge. I discuss these sets next in the context of theoretical and empirical advances since the publication of *Moral Minds*.

Features 1–4: the mature state of moral knowledge. Rawls's analogy to language was based, in part, on the apparent similarity between our grammaticality and aesthetic judgments, on the one hand, and our ethicality judgments, on the other. When we perceive an event, we spontaneously judge whether it is morally right or wrong, and we do so without access to the underlying principles. Though we may express a principle, it may or may not match up with the operative but unconscious principles driving our intuitive judgments. The analogy to aesthetics and, I would add, humor helps us think about this idea. Humor is a human universal. In every culture, people laugh at some things but not others. Usually, something that is funny is unexpected, unpredicted, and, when cast in the form of a joke, often entails a caricatured abuse of some party—think of all the

“dumb blonde” jokes or the spin-offs from “How many *Xs* does it take to *Y*?” When we laugh at a joke, we do so spontaneously. We don’t sit around and reflect on the details, although sometimes there are delayed reactions, and, in some academic circles, there are serious inquiries into what constitutes a good joke, including analyses of what tickles different parts of our brain. Jokes simply *are* or *are not* funny. Whichever way we lean, our judgment, expressed as a smile or laughter, is spontaneous, delivered without immediate access to the underlying process; whether one *could* access the principles is an open question, but given the lack of a coherent theory of humor, it seems more likely that the principles are abstract and inaccessible.

The fact that the various components or processes that enter into humor appear universal certainly does not rule out variation across cultures. When I was living in Kenya, my first movie experience in the capital of Nairobi was Dustin Hoffman’s *Tootsie*. Hoffman plays the role of a failed actor who suddenly finds success in playing the role of a woman on a daily soap opera. The twist is that no one on the show thinks that he is a man playing the role. And neither do the thousands of adoring female fans who find comfort in the strength of her—his—character. There were dozens of scenes that made me laugh out loud. What was bizarre is that I was the only one in the theater laughing and, I believe, the only non-Kenyan as well. The Kenyans did not think that a man trapped in a female role was funny. Who could argue with them? Imagine an American such as myself standing outside the theater trying to convince every exiting Kenyan that they had missed the point, and that it was really funny when Hoffman asked his friend whether he looked good in certain dresses and with particular kinds of makeup. You cannot convince someone that something is funny. Either they think it is or they don’t. In fact, *thinking* is not really the right description, as the process is so unthoughtful, so unconscious.

Saying that the principles underlying our sense of humor or morality are inaccessible does not imply that they are undiscoverable. The motivation for drawing an analogy to linguistics is to push for such discovery in the moral domain. Moreover, if such principles are discovered, and brought to the attention of every man and woman on the street, it is possible that they will directly impact upon our actions. It is also possible that they will not. The fact that linguists have discovered, and are thus aware of many computational principles, plays no role whatsoever in their day-to-day speech or writing, even if it enriches their appreciation for the beauty of what comes out of their mouth or pen. In this sense, the

language faculty is impenetrable, a modular system with only shallow outputs. Whether the moral faculty is defined by a similar architecture is, at present, unknown but certainly not unknowable.

The fact that some of our moral judgments are delivered without apparent access to the underlying principles takes the Kantian creature out of the main action and leaves the Humean and Rawlsian creatures to account for the intuitions. We are either morally dumbfounded in Haidt's terms because Humean emotions of disgust or empathy unconsciously drive judgments of permissible actions, or we are explanatorily challenged because we cannot access the principles of our universal moral grammar that determine how the causes and consequences of action figure into our moral decisions. Perhaps it is a bit of both. What I hope is clear from my droning repetition of this point is that the Humean creature requires the Rawlsian: no emotional response is possible in the absence of some appraisal system that evaluates the causes and consequences of action. And given my criticisms of the Humean sentimentalist position, I think it is fair to say that, whatever role our emotions play, they are unlikely to provide the essential ingredients for building the complex architecture that guides our moral judgments.

In some of Kahneman's experiments on fairness, subjects delivered highly consistent responses to changes in the market value of a commodity, with patterns that are consistent with prospect theory. They were, however, completely unaware of the reference-transaction parameter that alters our assessment of fairness. When we assess a transaction, our sense of fairness is mediated by an unconscious process that crunches through the profit and a set of reference terms for the individual or group in control of the commodity. This parameter was uncovered by Kahneman and his colleagues based on numerous experiments and hours of theoretical reflection.

As mentioned above in my discussion of the Mayan data, a culturally diverse group of subjects perceive means-based harms as morally worse than harms caused as a foreseen side effect. When subjects are asked to justify their judgments, few provide an explanation that appeals to this distinction. This suggests that the difference between means and side effects, a distinction that underpins Aquinas's doctrine of double effect, is operative in our moral judgments, but outside of our awareness. Recently, together with Philip Pettit and Bryce Huebner, we pushed further on the notion of *means*, asking whether there are situations where it is permissible to use someone as a means to a greater good. Based on the logic of

Pareto-improvement, we tested the idea, often discussed in philosophy, that it is permissible to use someone as a means as long as one does not make him or her worse off and the action makes others better off. Our first confirmation of this hypothesis emerged from a reanalysis of data collected by Greene and colleagues in their imaging study of subjects judging the appropriateness of actions in different moral dilemmas. Though the primary distinction made by these authors was between personal and impersonal dilemmas, the personal dilemmas could be further divided into Pareto and non-Pareto cases. For example, the footbridge-trolley dilemma is a classic case of a non-Pareto situation, as the fat man is used as a means to a greater end and is also made worse off (that is, he was merely an onlooker, minding his own business). In contrast, the crying-baby dilemma is a classic Pareto case: enemy soldiers will kill a mother, her baby, and the other townspeople hiding in a cellar if they are heard. As the soldiers approach, the baby starts to cry. If the mother smothers her baby, she and the townspeople will survive. If she allows the baby to cry, the soldiers will find and kill everyone. Here, smothering the baby does not make her worse off, but it makes everyone else better off—they survive. Significantly, and across a wide variety of dilemmas, subjects judged Pareto cases as more permissible (that is, endorsing the use of one as a means to a greater good) than non-Pareto cases and also took longer to respond to the Pareto cases.

To further push on the role of Pareto-improvement, we created several other dilemmas, and for each context (such as trolley problems, a burning house, manipulated both the extent to which the person used as a means was made worse off and the extent of physical contact with this person. Consistently, subjects judged cases that satisfied Pareto-improvement as more permissible than non-Pareto cases, and judged cases involving physical contact with the individual as less permissible than in the absence of physical contact (for example, throwing a rock). To illustrate, consider a trolley case in which there is only one track, one person near the oncoming trolley, and five people farther up ahead; if the trolley continues, it first kills the one, and then the five. If a bystander throws a rock at the one, he screams, but the five hear this and get off the tracks in time. Ninety percent of subjects endorsed throwing the rock even though this made the one person worse off. In contrast, only 65 percent of subjects endorsed throwing a dying man onto the tracks to save five up ahead; here, there is physical contact, *and* the dying man is made somewhat worse off relative to the one man on the tracks.

Turning back to the issue of accessibility, is a parameter such as the *reference transaction* or *Pareto-improvement* part of our universal moral grammar? I doubt it. My guess is that these factors, though important for moral decisions, are neither sufficiently abstract nor sufficiently detailed to explain the richness of our cooperative and harmful interactions, including those aspects that cohere across cultures and those that are open to some level of cross-cultural variation. Though we are not yet in the position to offer principles or computations with greater explanatory depth, we can be relatively confident that those on offer thus far are insufficient. For example, though philosophical analysis has been thrown at the family of trolley dilemmas for more than thirty years, neither the double effect nor Pareto-improvement has the explanatory breadth to explain the variety of cases where harm is either permissible or not. To achieve this level of analysis will require a completely new way of looking at the problem, and, minimally, a recognition of three distinct components of the moral faculty: the computational resources that enable infinite expressive power, a set of dedicated concepts, and a sensory-motor system that represents our actions. Some of these components may be unique to morality, some shared with other domains of knowledge. It is also possible, perhaps even most likely, that none of the components on their own are unique to the moral faculty. Instead, what is unique is how each component interfaces with the others to create morally specific outputs. This is a line that the post-Rawlsian moral grammarians have pushed, and I believe it is on the right track for both morality and other domains of knowledge.

Recently, I explored the issue of interfaces with Huebner and several students. In particular, we were interested in how the utilitarian calculus that evaluates the greater good interfaces with the systems of number representation. As mentioned in Lecture I, a great deal of work over thirty years has established that nonlinguistic animals, prelinguistic humans, and fully competent linguistic adults have access to two nonlinguistic systems of number representation. One system computes number approximately, constrained by Weber ratios, and based on representations of analog magnitudes. A second system computes number precisely, is limited to counts of up to about four, and is based on representations of individuals, in what has often been called the system of parallel individuation. In a first series of experiments, we took the standard bystander trolley problem and explored different numbers of people on the main and side track and asked, using a 7-point permissibility scale (7 = forbidden, 4 = permissible, 1 = obligatory or required), for a judgment about flipping the trolley onto the

side track, toward the smaller number of individuals. Overall, regardless of the contrast, subjects judged as permissible all actions in which at least one additional individual was saved. Thus, for example, subjects stated that it was permissible to turn the trolley onto 1 versus 5, 2 versus 5, 3 versus 5, 4 versus 5, 1 versus 2, 1 versus 3, and 1 versus 4. But it was also permissible to turn the trolley for 100 versus 500, 100 versus 200, and 101 versus 505. Perhaps more strikingly, when we asked an open-ended question about how many people would have to be on the main track before it was *obligatory* to turn the trolley onto the side track with one person, the modal answer was 2! Paralleling George Miller's famous magic number 7 plus or minus 2, the utilitarian calculus evaluating harms is mediated by a simple logic and the magic number +1. That is, as long as one additional life is saved, it is permissible to turn the switch, and, for some, obligatory. Thus, though some aspect of these number systems interfaces with our moral judgments, the utilitarian calculus is unconstrained by their operation.

Raising the analogy to linguistics, and invoking the possibility of a universal moral grammar, forces us to ask different questions about the nature of our moral judgments. For example, is it possible that in the generation of an "obligatory" verdict, all moral dilemmas look the same at some abstract level, even though some scenarios target harm whereas others target help? What matters in the moral calculus is whether the net benefits are positive for some recipient relative to the costs to the agent. Is the distinction between *harming* and *not helping* an artifact of language, with the moral faculty myopically focused on avoiding negative consequences? When we confront the opportunity of rescuing a drowning baby, at no personal cost, we feel the moral obligation to do good—to save the baby. But our moral faculty may see it the other way around: it is morally obligatory because *not rescuing* the baby—omitting the action—has negative consequences. The moral pull comes from the negative consequences, not the virtues associated with doing good. What we need is a way to formally describe the causes and consequences of actions within an event, including their temporal order and the possibility that our judgments arise from operations that are not visible at the surface. Let's make this more concrete by going back to the bystander version of the trolley problem. Though it may seem that I have worked this example to death, we can go deeper by breaking the event down into finer components, organized as a function of time, with each piece of the event handed off to the systems involved in computation, conceptual analysis, and sensory-motor representation. As I mentioned, although the price of focusing myopically on a small corner

of our moral psychology is that we are missing out on the richness of this domain of knowledge, the benefit is that we gain increasing explanatory depth and open up new questions.

Every event entails a sequence of causes, actions, and consequences. The computational system consists of rules that enable different combinations of these elements, in particular orders and with specific relationships. The computational system provides the syntax for how these elements relate to one another. The system that is involved in conceptual analysis (handling issues of agency, intentionality, and goal) represents the event in a fundamentally different way than the system that represents the physical actions and consequences (handling issues of contacting, holding, and moving). Yet, to derive a moral judgment, these systems must *talk* to each other. Something in our brains must make the information from these systems mutually intelligible. How this works is a genuine puzzle. Recognizing the fact that there is a puzzle—one worthy of a solution—represents the first step.

Consider the bystander case, again. The bystander or agent first perceives a trolley out of control and five people (animate objects, recognized as such by a system that is involved in moral judgments but not specific to it) ahead in harm's way. Recognizing that they are in harm's way constitutes a projection that the target or object is of a kind that can be harmed. It also entails, if this is a real-life event, a prediction of the trajectory of the trolley, presumably by the system that handles problems of naive physics. The agent can see that if the trolley continues it will run over and kill the five people; the agent perceives physical cause and effect, another capacity that is not specific to morality but enters into processing. At this stage of processing, there is no moral dilemma. There are no competing options. Once we introduce the switch, its function, and potential consequence, a dilemma arises. The agent sees the switch (inanimate object) and a person (animate) on the side track, out of harm's way. The agent understands that the switch can physically cause the trolley to switch tracks. The switch provides the means to transform each person's psychological state from living to dying or the opposite. Leaving the switch alone results in five dying and one surviving; flipping the switch results in one dying and five surviving. The act of switching carries no moral weight of its own. It is represented by the sensory-motor system as an action that involves using the hand to displace an object from one location to another. To generate greater variation, and add to abstraction, it is possible that the mind represents the action SWITCH as \pm SWITCH, where + stands for *act* and – stands for *omit*.

In this way, action/omission counts as a parameter, one that is bound to every possible action imaginable. From this account, we can translate this piece of the event as:

BYSTANDER \rightarrow \pm SWITCH

In more general terms we have:

AGENT \rightarrow \pm ACTION

The physical act of pulling the switch (+SWITCH) predictably leads to particular consequences, with differences in timing. In this sense, each action projects to different consequences with respect to some target object or objects. Flipping the switch immediately results in five surviving and then, with some delay, leads to the trolley switching tracks, contacting the one person, and then the one person dies. Again, the observer or judge passes part of this event to the sensory-motor system that represents CONTACT and the other part to the conceptual system that represents SAVE and DIE. All of this happens, abstractly, as the agent sets up different possible outcomes for this scenario. With these details in play, the agent creates a goal with an intended set of actions and a recognized set of outcomes. This involves additional concepts, including NUMBER (both the specific values and their ordinal relations), INTENDED, and FORESEEN; recall that in the switch case, the agent's PRIMARY GOAL is to save the five people and, as a FORESEEN consequence, kills one person. Like language, it is possible that the moral faculty unconsciously *displaces* the consequences of the switching action to the front of the decision process, such that the motivation for moving forward with one choice is the utilitarian outcome of favoring more lives saved over fewer. In other words, although the actual, physical consequence of the action happens downstream, the agent's representation may bring it up front in order to advance the decision or permissibility judgment. Using language to express this thought, we would have: five saved is better than one harmed, so flip the switch. Whether the moral faculty actually moves particular pieces of an event around in order to run the computation is pure speculation at this point. What is critical is to appreciate that each event has different elements, that these elements are related, and that the moral faculty's analysis of the event may involve reordering the actual timing of elements within the event. Although consequences naturally flow from causes, the moral faculty may invert this process, looking first to consequences and then to causes.

Our conceptual system handles the intention and the nature of the agent's goals. The part of the sensory-motor system that focuses on action handles the internal calculation of motoric organization entailed in moving a switch. Whether there is anything like a set of distinctive features for action—what we might consider as *actemes*—that parallel those uncovered for phonology is unknown, but certainly possible, and a topic that has long been discussed in research on motor programming. In the case depicted, the agent first intends to flip the switch. For each action, the conceptual system may bind a further feature: Intended (I) versus Accidental (A). From this perspective, we can rewrite the bystander case as:

$$\text{BYSTANDER} \rightarrow (+I, -A) + \text{SWITCH}$$

Here, the bracketed portion stands for an intended act (+I), but not an accidental one (-A). And as before, we can make this more general by substituting AGENT for BYSTANDER and ACT for SWITCH. And we can also add that the consequences are foreseen (+F) yielding: AGENT \rightarrow (+I, +F) +ACT.

Importantly, however, there is not one foreseen consequence, but at least three relevant ones: the trolley switches tracks, the trolley moves away from and avoids contacting five people, and the trolley contacts and kills one person. As Mikhail and Sinnott-Armstrong have noted, these events are also time stamped, and, importantly, the timing matters as the trolley moving away from the five is tagged as a *good*, whereas the trolley moving toward the one is a *bad*. Since the good happens before the bad, the case tilts toward permissibility. And if timing matters, as it does in people's folk intuitions, then if you play with the order of events such that the bad happens before the good, subjects judge the case as less permissible than in the opposite situation.

The consequences of an event are also processed by the sensory-motor and conceptual systems. On the sensory-motor side, something like the mirror-neuron system may engage, allowing an individual to both act as an agent or experience what it would be like to produce these actions when someone else is responsible. On the conceptual side, the first and second consequences are represented as intended, whereas the third is represented as foreseen (F). Following the description above, we write (+I, -F) for a consequence that is intended (+I) but not foreseen (-F) and (-I, +F) for an act that is not intended but foreseen. We can now build on our principle above:

BYSTANDER $\rightarrow (+I, -A) +$
 SWITCH $\rightarrow (+I, -F)$ TROLLEY MOVES $\left\{ \begin{array}{l} \rightarrow (-I, +F) \text{ 1 PERSON DEAD} \\ \rightarrow (+I, -F) \text{ 5 PEOPLE SAVED} \end{array} \right.$

The conceptual system assigns roles to different components of the event, including perhaps something equivalent to subject, verb, and object, with temporal markers to indicate when actions and consequences occur. Given that both actions and consequences can be intended, some part of the system must make this distinction or the interpretation will crash. The computational system provides structure for the ordering, combination, and relationship between these elements.

By accepting the minimal characterization of the moral faculty as a capacity with three distinct components, we also implicitly accept the fact that each component must interface with the others. To repeat, something, somewhere in the brain, must be involved in the translation. Something must make the output of one system legible to the other, given that each system represents information in fundamentally different ways. The action system presumably stores something like a library of gestures for articulating our body, and for perceiving these gestures. The mirror-neuron system (perhaps among many others) assumes some responsibility of unifying action and perception, presumably into one neural code; several researchers working with cortical areas housing mirror neurons suggest that the coding is at the level of subtle gestural differences, perhaps hinting at something like a set of distinctive features. But this system must then pass on information, or receive information, from the conceptual system that sets up the action's goal and the agent's intentions and, presumably, predicts the potential array of consequences, both good and bad, for self and other. These are different representational formats, and, somehow, they must talk to one another. How this happens is unknown, with many of the same puzzles facing those working on language.

I realize, and to some extent apologize for, the abstractness of this discussion. But to repeat the common refrain: our current descriptive principles are insufficient for the moral domain. They neither account for what appears to be universal among moral systems nor account for the variation between systems. The analogy to language suggests that we look elsewhere. In particular, we need to understand how the moral faculty enables a limitless range of moral actions and their comprehension, how each child acquires this system from a presumably impoverished environment,

and why no other animal does. Our best guess thus far is that there is a set of computations that enable unbounded expressions based on representations of actions and concepts of cause and consequence. The linguistic analogy also suggests that we will have to abandon the commonsense terms used to describe moral judgments. For example, in the same way that linguists have abandoned discussion of *passive constructions* (for example, “The motor was fixed by Fred”) for more elementary and discrete computations or processes (such as morphological changes [add -ed], an operation based on the thematic role of an element such as *agent* [Fred] and movement of the object [*motor* moves up front in the sentence]), the same may be necessary to describe why different events are judged in different ways. Saying that an event is *fair* is not enough. We need to know how different components of the event interact to generate this judgment, while entertaining the possibility that superficially different events may be guided by similar operating principles. Achieving this level of description is really a necessary first step before proceeding to the problems of moral development and evolution.

Features 5–7: acquiring a mature state of moral knowledge. One reason to be optimistic about the principles and parameters approach to morality that I defended in *Moral Minds* is that it makes the acquisition problem more manageable in the face of variation in moral judgments. As detailed studies of language reveal, in the absence of fixed principles to guide acquisition, the child faces too many options for generating a phrase and interpreting someone else’s. With the theoretical introduction of parameters, it was possible for the first time to see a solution to the problem of language acquisition: how the child goes from some initial state to the mature state that entails a set of rich, internalized principles that can generate and comprehend an unbounded number of grammatical utterances in the native language. At each developmental turn, the incoming linguistic data sets the relevant parameters, guiding the child toward its native language. Of great interest at present is whether any of the parameters start out with default settings, and how the child “decides” which properties of the input should count as evidence for setting a parameter. Looking back to Lecture I, this view of language sees the role of the environment as a selective agent, whereby the options are all given. Guided by this framework, linguists have made great strides in understanding how each child grows its native language. Note: this view of the process admits of both nature and nurture, to trot out that old and tired dichotomy. The incoming linguistic

data set the parameters and put the child on its path to growing a specific language. This is nurture, but the parameters themselves are part of nature. Further, because the linguistic data are insufficient to account for the mature state of knowledge, we invoke the principles that constitute our biologically endowed universal grammar.

Drawing on the analogy, the key to understanding moral development is to recognize the problem: based on a limited amount of time, as well as a limited and impoverished input, each child acquires a richly textured system of moral knowledge. By invoking a fixed set of principles as part of the child's universal moral grammar, together with a set of parameters, the child's path to acquisition is guided to a highly predictable end point. Growing a particular moral system reduces to setting parameters. The open challenge to those like Piaget and Kohlberg who wish to see this system constructed from experience is to show both that the input is richly informative and that it can carry out the required tutorial, guiding the child from morally ignorant to morally sophisticated. Thus far, those following in this empiricist tradition have not delivered on their account. Part of the reason for this is that they have focused almost entirely on what the child expresses when she acts in a morally appropriate or inappropriate fashion, leaving her moral competence completely unexplored. And even at the level of expressed explanations for morally permissible actions, it is not clear how they acquire this knowledge, moving far beyond the input, both casually observed and explicitly handed down from parents, teachers, and other wise elders.

Even the most ardent empiricist has to grant some innate structure. Though dogs and cats are raised in the same environment as human children, they are neither appalled by our moral infractions nor delighted at our virtuous acts. Of course, Fido may feel the weight of being scolded when he sneaks out food from the refrigerator or glee when he is rewarded for barking at an intruder, but these emotions do not constitute moral evaluations of the situation; at least we have no reason to think so. Something about us, but not them, makes us different. What that difference is—what the initial state looks like—is anyone's guess at this stage. In the absence of rich descriptive principles, it is difficult to formulate a precise set of developmental questions. My approach, however, has been to assume that something like the principles and parameters view is correct, with the moral faculty operating over the causes and consequences of action.

In *Moral Minds*, I explored two of the child's early abilities: the capacity to generate expectations about the physical world, detecting consistencies as well as violations, and to use the psychological attributes of objects to infer goals. This search uncovered a surprising degree of sophistication. Before their first birthday, infants generate expectations about physical events, looking longer at impossible movements, occlusions, displacements, and appearances. Although they do not have the linguistic capacity to label these violations, the contrast between consistent and possible events with inconsistent and impossible events may form the basis for an early, purely descriptive system of right and wrong. This might be accomplished by marrying the capacity to detect violations with the building blocks of action comprehension that have been uncovered by developmental psychologists such as Leslie, Baillargeon, Bloom, Carey, Csibra, Gelman, Gergely, Keil, Premack, Spelke, and many others. Although none of the principles guiding action perception and analysis in infancy are specific to the moral domain, they are presumably necessary. Normally developing infants distinguish animate from inanimate objects, use contingency and environmental flexibility to infer goals and agency, and combine several of these inferences to classify interactions as positive or negative, worthy of avoidance or continued engagement. Like dogs and cats, infants may not be appalled when a circle fails to aid a triangle against a bigger and *tougher* square, and may not applaud an amorphous blob for picking up another equally amorphous blob drowning in a pool of water. But if they detect inconsistencies as revealed by their looks, and are sensitive to goals and contingent interactions, then the starting state minimally consists of these folk psychological competencies. How they connect up with the system that distinguishes moral from other social dilemmas, and ultimately adjudicates on the moral permissibility of an action, is currently unclear.

There is another reason to take the nativist position seriously, and especially the idea that each child is equipped with a universal moral grammar that both constrains the range of possible moral systems acquired and, at some level, cordons off impossible moral systems either because they are unlearnable or because, if learned, they would be unstable; as a refresher, this is the moral space referred to earlier in Lecture I. The reason is simple, derives from Chomsky's discussion of meaning for natural language, and is framed here as a question: why is it that we *do not* acquire certain interpretations, perceptions, actions, or responses? Why are certain moral interpretations licensed but not others? The philosopher of language, Paul

Pietroski, states the point clearly in discussing different theories of meaning: “Facts about how humans *don’t* associate signals with interpretations may well reveal important aspects of *how* humans understand language—especially if such facts raise theoretically interesting questions about how children manage to converge (in so far as they do converge) on agreement about signal-interpretation associations, despite disparate and often relatively impoverished experience.” For language, the idea is that something must constrain how we construct interpretations for different expressions, and why the art of interpretation is not more chaotic, open to local cultural fads, the time of day, or weather. To illustrate, consider one of Pietroski’s examples. Native English speakers readily see that sentences (2) and (3) are plausible interpretations of (1), but sentence (4) is not:

- (1) The millionaire called the senator from Texas.
- (2) The millionaire called the senator, and the senator is from Texas.
- (3) The millionaire called the senator, and the call was from Texas.
- (4) The millionaire called the senator, and the millionaire is from Texas.

Children raised in an English-speaking environment also derive these interpretations even though no one told them that (4) is blocked as an interpretation whereas (2) and (3) are permitted. It is not that ambiguity *per se* is blocked in language—ambiguity is rampant in spoken language! It is also not the case that blocked sentences (4) are incoherent or the culmination of conceptual word salad by a linguist attempting to foil the child’s attempt at understanding. So why certain interpretations *do not* follow from what we hear is as valid a problem as why certain interpretations *do* follow.

The same logic applies to the moral domain. We want to understand not only how and why children assign moral weight to particular actions but why they do not assign such weights to everything they perceive. Saying that children readily perceive an action as a social convention, and not a moral rule, does not explain why they perceive things this way. It merely redescribes the act. Saying that moral norms are prescriptions for what we ought to do together with a punitive chart for infractions also does not help. It only redescribes what is associated with an act, as opposed to the principles that determine its status. As I alluded to earlier, philosophers Daniel Kelly and Steve Stich have argued that the distinction between social conventions and moral rules may turn out to be pure mythology, an artifact of presenting paradigmatic cases that fit as opposed to exploring the broader landscape that, upon inspection, does not fit. What we

want to understand are the principles that cause certain actions and their consequences to achieve moral status, and how children within a culture eventually converge on the same moral interpretations despite variable experiences.

Part of the moral-development problem has strong parallels with the sensory-motor aspects of language acquisition, and in particular our phonological representations. How, for example, does the child, bombarded with environmental sounds, pick out speech from the grumbling of a vacuum cleaner? How does she distinguish speech and the other sounds that come out of human mouths, including laughter, hiccups, screams, and, toughest of all, the *ums* that interrupt most fluent discourse? Work in psycholinguistics reveals an early sensitivity to speech as opposed to other sounds, including a listening preference for speech over other matched sounds,¹⁸ an ability to discriminate phonemes in utero, and a left-hemisphere processing bias for speech but not other sounds. It seems that the child is born with a universal inventory of distinctive features or phonemes, with the initial set whittled down by the local culture. Whittling is probably not the right metaphor, as those phonemes that lack employment in the native language are still represented in the brain. More appropriately, the process of environmental selection picks out and emphasizes the relevant phonemic contrasts, while suppressing (not rejecting) the irrelevant ones. Linguist Charles Yang has suggested a similar process for other aspects of language development, including the implementation of particular principles and parameters.

We might imagine a similar process operating in the moral domain. The child is endowed with an inventory of actemes—each discrete, meaningless, and potentially combinable with other actemes. Interactions with the environment set up which actemes join to others, generating an output that can be interpreted as meaningful by the conceptual system. The system that represents action also interfaces with the combinatorial machinery to create a limitless array of action sequences, and, also, to impose hierarchical structuring on these sequences; for example, actions combine to create subgoals, which combine to create higher-order goals, which combine to

18. Even here we must be careful! Recent work that I have carried out in collaboration with Athena Vouloumanos, Janet Werker, and Amelia Martin (A. Vouloumanos, M. D. Hauser, J. F. Werker, and A. Martin, "The Tuning of Human Neonates' Preference for Speech," *Child Development* [in press]) reveals that the early preference for human speech may, in fact, be an early preference for primate sounds or, perhaps more generally, mammalian sounds. Thus, neonates less than forty-eight hours old will suck a nonnutritive nipple for as long to playbacks of human speech as they will for rhesus monkey vocalizations.

create an event. If anything like this is going on, then it sets up a string of additional questions, once again paralleling language: Once a child acquires its suite of distinctive action features, are those implemented in a different moral system as incomprehensible as the “*r* to *l*” distinction in English is to native Japanese speakers? How does the child break into this system, segmenting the continuous stream of actions within an event into functionally meaningful units? We know from recent studies of speech that human infants are endowed with a set of basic statistical abilities that enable them to track the distribution of phonemes in the native language. By picking up on the native distribution, the statistics provide one way to segment the incoming speech stream. Notably, this mechanism does not appear to be specific to speech, raising the possibility that it could be deployed in the service of action analysis in the moral domain; Dare Baldwin’s research on how infants parse a complicated event into action segments related by statistical properties provides some suggestive support for this idea.

Other questions arise in terms of how the child makes use of input in setting certain parameters and, thus, fixing the moral judgments assigned to principles such as those that underlie fairness, harming, and helping. When we tutor our children about the moral rights and wrongs of our society, is this as ineffective as our trying to correct their grammatical boo-boos (It’s not “*I go-ed* but *went* to the market”)? When our tutorials—with carrots or sticks—are effective, is it because we are correcting mere superficialities of the system, akin to the schoolmarm’s usage handbook? Is it like correcting etiquette as opposed to setting up the inaccessible principles and parameters? When parents label actions as right or wrong, what does the child learn? One possibility is that this kind of input shapes the child’s culturally specific moral knowledge in the same way that the input from her native language shapes her lexicon. But what about all the things children do that never receive comment? For Kohlberg, children at the early stages of moral development are guided by the explicitly held belief that actions that are bad are punished whereas actions that are good are rewarded; if this were the correct view, then children ought to conclude that any action that is not punished is good. But they do not draw this conclusion. Concepts such as good and bad, permissible and forbidden, are abstract, triggered by an ungodly variety of actions with little to no superficial similarity: hitting brother and a delicate vase, putting food in the mouth or up the nose, spitting out toothpaste into the sink as opposed to on top of someone’s head. Yet early in development, children apply these

concepts in the appropriate context and, well before this, comprehend their own or another's actions in such normative terms. This suggests that the child may be innately endowed with such concepts, primitives of the moral faculty. As in the case of our action system, what we want to understand is how the child's variegated experiences map onto these concepts, and the extent to which there is plasticity in the system.

Consider the economist's ultimatum game and, especially, the cross-cultural data collected from several small-scale societies. In brief, the classic ultimatum game involves two players, a donor and receiver. The game is played once, anonymously. Both know the rules. The donor starts with ten dollars from the bank, and is allowed to give some proportion to the receiver; if the receiver accepts, the donated amount is given away, leaving the remains for the donor; if the receiver rejects, neither the donor nor the receiver obtains any money.

Given evidence that all cultures have some notion of fairness but impose different limits on the permissible offers and rejections, we can ask two questions: what experiences set each culture's permissible range, and, if individuals from one culture migrated into another, what, if anything, would cause their settings to change? If the settings changed, would the acquisition process be similar to that of their first or native system? Concretely, the Hadza—a group of Tanzanian hunter-gatherers—tend to donate extremely low offers, whereas the Ache hunter-gatherers of Paraguay offer close to half. Following emigration to an Ache group, what would it take to shift the Hadza's donations? Sanctions would presumably cause a shift in performance, but would their competence change as well? Would the process of change be long and arduous, requiring tutelage and punishment? Or would the change be as simple and trivial as driving on the opposite side of the road—the American experience of driving in England?

The developmental questions naturally bleed into questions concerning the specificity of the moral domain. Are there mechanisms specially dedicated to solving moral problems, systems that enable the child to immediately pick out moral from nonmoral events? Or is the moral arena the consequence of combining domain-general processes? In discussions of domain specificity, patient populations are illuminating. Consider the process of categorization. Many have assumed that this is a domain-general process that is operative for faces, food, cars, houses, colors, language, music, and, yes, morality. When patients show category-specific deficits, with one category damaged and the others spared, the domain-general perspective has been injured. Damage to a small bit of the temporal lobe causes a

loss of face recognition, but no other recognition abilities. This bit of cortex must be dedicated to face processing. Some patients cannot recognize fruits and vegetables but are fine with other foods as well as nonfoods, suggesting that the mind is endowed with something like a produce module. Some patients have trouble recognizing vowels but not consonants, as well as the reverse. Domain-general theories cannot explain the nature of these deficits. Is there anything comparable in the moral domain?

In parallel with the other developmental issues raised here, we are on uncertain footing until we have a richer description of the mature state of moral knowledge and its guiding principles. That said, there is some suggestive evidence concerning the selectivity of the moral faculty. Recent work by Cosmides, Tooby, and colleagues on the Wason selection task is of relevance. Although we readily acknowledge the category of permission rules, there appear to be domain-specific mechanisms for handling the details of different rules, including social contracts and precautions. Early in development, children appear sensitive to the fact that there is not a single generic social rule but rather a suite of distinctive social rules. For example, at the age of three to four years, children distinguish between a *deontic rule* that indicates that an act *must* or *should* be done (“All noisy children must play inside the house”) and an *indicative rule* that makes a claim about the current state of affairs (“All noisy children are playing inside the house”). To make this distinction, children must be sensitive to such morally loaded words as “must,” “should,” or “ought.” In the deontic case, they are looking for individuals who violate the rule, whereas in the indicative case they are looking for confirming evidence. Perhaps the strongest evidence in favor of the domain-specific position comes from a patient who shows normal competence with precautions but a severe deficit with respect to social contracts. If there were a domain-general system for processing rules, such patients would not exist.

The work on ventromedial prefrontal cortex patients, reviewed earlier, provides another case. As noted, these patients show a highly selective deficit with respect to moral judgments. In particular, if we consider the moral space, they show normal patterns of judgment for impersonal moral dilemmas, as well as for personal dilemmas in which the act is completely self-serving. Where the deficit arises is in the context of personal dilemmas that put into conflict a highly aversive and harmful act against a highly beneficial and helpful act. In these situations, the patients lean in the direction of consequences, showing a form of hyperutilitarianism. Intriguingly, preliminary work by Shriver suggests that patients with damage

to the dorsolateral prefrontal cortex present with the opposite pattern: that is, they appear to ignore the consequences and state that, if the act is harmful, it is forbidden. Studies such as these, combined with both neuroimaging experiments and studies using transcranial magnetic stimulation to deactivate particular parts of the cortex, are likely to begin unpacking the complicated circuitry underlying our moral judgments. That said, we should perceive these advances cautiously, placing them in the context of our relatively poor understanding of the neural basis of language, a domain for which we understand far more.

Features 8–10: evolving a mature state of moral knowledge. In discussing the evolution of morality as a domain of knowledge, the issues that arise are to some extent specific to morality and to some extent quite general. On the general side, whenever we consider the evolution of a domain of knowledge, we want to ask whether it is both unique to a given species as well as unique to the domain. Failing to make this distinction can lead to senseless debates. Here, therefore, I want to ask whether some of the processes that humans recruit in the moral domain are uniquely human. We can answer this question only by adopting a broad comparative approach (sampling a variety of animals) as well as a broad ontological approach (sampling a variety of domains). If we want to stake out the claim that a mechanism is uniquely human and unique to a given domain of knowledge, then we need to take the following steps. To address the issue of uniqueness, we need to look at species other than humans. If we are strictly interested in the possibility that this mechanism evolved by descent from a common ancestor, and constitutes a homology, then our best bet is to look at the other primates given their evolutionary proximity. Finding evidence of this mechanism in primates rules out the claim that it is uniquely human but also leaves open the possibility that it evolved before the primates, either once or independently in different animal lineages. If we fail to find evidence of this mechanism in the primates, this should not be the end of our comparative search. It may be that other species, more distantly related, confront problems that are much more similar to our own. For example, due to the hierarchical arrangement of notes into syllables, and syllables into song, that we observe in some birds and whales, it may be that these species will exhibit greater parallels with humans when it comes to some of the essential computational components of our language faculty. If, on the other hand, we find no evidence at all that the target mechanism is present in species other than humans, then the uniqueness claim holds, but the uniqueness of the domain remains open for challenge. Test-

ing whether a particular mechanism is unique to a domain requires tests that move across domains. For example, and as mentioned in Lecture I, both language and music exhibit hierarchical structure, and both use a combinatorial operation to create an unbounded level of expression from a finite set of elements. At this level of detail, language and music share similar resources, ruling out the idea that these mechanisms are unique to one domain of knowledge.

There is another set of evolutionary questions that we must add: What is a particular domain of knowledge for? What adaptive problem does it solve, and to what extent do its design features reflect a history of natural selection together with other random and nonrandom evolutionary forces? To what extent do certain mechanisms that enable this domain reflect the perhaps inevitable outcome of certain design constraints? For example, given the physics of nerve cells and the circulatory system, it is not possible to innervate or send blood to something approximating a wheel, even though a number of terrestrial animals would benefit from this transportational advance. Returning to language, it has often been argued that language is “designed for communication.” When we look at the design features, we see a system that was as exquisitely designed for communication as the eye is designed for seeing. Answering questions of the “What is it for?” form is notoriously difficult as they inevitably force a distinction between original and current functions. No one can debate the fact that we use aspects of our language faculty to communicate, nor can one debate the plausibility of the idea that this faculty evolved to solve the problem of communication. But as discussion of its computations reveals, the faculty of language is involved in other aspects of our mental life, some of which are never expressed. The faculty of language also enables us to organize our thoughts, plan for the future, create new conceptual resources by combining and recombining more primitive concepts, and so on. Our language faculty is used for all of these things. A more sensible question might therefore be to ask what different components of the language faculty are for, leaving the complicated ways in which they interface for a later date. Regardless of how this debate turns out, my goal here is merely to flag the challenges as we turn to morality.

Like human infants, nonhuman animals set up expectations about physical and psychological events and look longer when an unexpected event arises, often because they have detected a principled violation. Though individuals are presumably not aware of these principles—they do not reflect upon the events *in order* to deduce *that* a violation has

occurred—they may form the foundation for making judgments about right and wrong. Like infants, at least some animals are equipped with a set of principles that guide their analyses of actions, both their causes and their consequences. Though work on animals has only just begun, my hunch is that we will share these primitive aspects of event perception. When it comes to segmenting a continuous stream of motion from an event into a discrete set of actions, perhaps hierarchically organized, there is increasing evidence that we are not unique. For example, initial studies by Call and colleagues showed that, like young infants, chimpanzees make a fundamental distinction between an actor who is *unwilling* to give food and an actor who is *unable* to do so. This distinction, subsequently demonstrated in monkeys as well, is important with respect to the notion of moral building blocks as it sets up a core difference between the *means* by which we achieve a target goal and the goal or *outcome* itself. Had nonhuman primates only focused on the outcomes, they would represent a phylogenetic snapshot of our early ontogeny, with young children fixated on consequences and only later appreciating the significance of integrating consequences with means to attain a more substantive moral appreciation of social events.

Studies of primates, and other animals as well, have also begun to explore other aspects of event perception that figure into our moral calculus, though, to be clear, in none of these cases are we licensed to conclude that the capacities are sufficient for moral agency. For example, studies of apes, monkeys, and dogs reveal that, like human infants, these animals take into account environmental constraints and a notion of efficiency in evaluating the rationality of an action. As one illustration, experiments reveal that both monkeys and apes perceive the rationality of an agent who uses his elbow to communicate a concealed goal when his hands are occupied (environmental constraint) but perceive this very same elbow gesture as irrational (non-goal directed) when one hand is free and *could* have been used to indicate the target goal. Further, several recent studies of imitation in chimpanzees reveal that they not only attend to the minutiae of a complicated action sequence but also recognize goals and subgoals, using this information to re-create an event in the service of creating social traditions.

A number of comparative issues remain wide open in terms of the uniqueness challenge, a point raised in Part 1. For example, our current understanding of animal emotion is thin at best. We can say that for some of the basic emotions, such as fear and anger, we share with other

animals some of the core behavioral and physiological signatures. For the more morally specific emotions, including guilt, shame, envy, empathy, sympathy, and awe, much less is known. If we use as a guideline the context-specific expressions that humans use to convey particular emotions, behavioral observations of especially primates and dogs suggest that something similar may be going on. But detailed analyses in other areas indicate that behavior can be a misleading guide. For example, although some of the moral emotions may operate, especially early in development, without a sense of self or other, sometime after the child's fifth birthday, emotions such as empathy, guilt, and shame take on a different complexion. Guilt sits within the broader context of what others believe. To understand what others believe requires a theory of mind, a capacity that shows significant maturation around the fifth birthday. Though studies of dogs and chimpanzees are beginning to show that these animals are endowed with some of the rudimentary properties of this ability—for example, some level of goal and intentional attribution—it is not yet clear how far this work will push. Should it turn out that animals lack a rich sense of self and other in terms of propositional states of belief and false belief, it would greatly limit the richness of the moral emotions, if they even have them. And if this system is impoverished, then important details of the Humean creature are missing, having evolved uniquely within our own species.

A primary reason for the gap in current understanding is that work on animals has focused almost entirely on interpretations of what animals do, as opposed to how they perceive and possibly judge what others do or might do. With the exception of a few studies, almost all of the work that is relevant to morality targets how animals adhere to social rules and what happens when there are infractions. But if the competence-performance distinction holds for animal minds as well, and it seems to me that there is no good reason to reject this possibility, then we will need an appropriate set of tests.

All socially living animals function according to a set of rules or principles that at least implicitly determine what is allowable. Paralleling some of our own principles, animals do not adhere to the deontological rule that *killing is wrong*. Rather, they adhere to principles of harm that are open to parametric variation. In some species, siblicide is obligatory. In other species, it is facultative, depending upon the vagaries of the local ecology. Within a variety of species, infanticide is allowed, sometimes the responsibility of the mother, at other times the charge of the new male

in the group. Rules for hierarchical structure are also variable and, again, influenced by the ecology as well as details of the mating system. In some species, rank is inherited, passed down from mother to daughter, generation after generation. Males reap this benefit or burden as long as they live within the natal group, but things change once they emigrate out; in some species, males fall to the bottom of the hierarchy as soon as they change groups, whereas in others they rise to the top. In understanding these social rules, we wish to understand not only what causes variation within and between species but how animals perceive violations. These will not be easy experiments, but they must at least be put on the table if we are to engage, at a comparative level, with the possibility that animals are endowed with a certain level of moral competence that may not align with their moral behavior.

Some of the neural circuitry underlying our moral behavior, especially those aspects that are supportive of as opposed to being restricted to the moral domain, is shared with other animals. Thus, we see evidence of mirror neurons, circuits dedicated to inhibitory control, processes involved in conflict resolution, emotional expression and processing, action perception, and social cognition. Some of these noted similarities are superficial, relying on coarse analyses. This is in part due to the fact that while recording neural activity from single cells, experimenters are necessarily forced to present fairly simplistic and often artificial behavioral or perceptual tasks. We are therefore left with only loose descriptions of how particular parts of the brain do or do not serve similar functions.

Though many of the interesting facets of our moral knowledge have yet to be explored in animals, some of those that have suggest that we are uniquely endowed. Consider cooperation. Though a wide variety of animals cooperate, it looks as though reciprocal altruism may be a uniquely human form of cooperation. More than thirty years' worth of research, dating back to Trivers's classic paper on the problem, has failed to come up with a single convincing case. Some of the examples—including de Waal's work on chimpanzees and capuchins, my own work on tamarins, and Milinsky's work on stickleback fish—come close, but for a variety of reasons fall short, and in interesting ways. The upshot is that animals lack some of the critical ingredients that enter into reciprocity and enable humans to uniquely stabilize their reciprocal relationships. These include the capacities for temporal discounting (delaying future rewards), the detection of cheaters, and punishment. Though human reciprocity can and does break down, our potential to maintain reciprocal relationships

is unmatched. Moreover, we are uniquely able to sustain large-scale cooperation with unrelated individuals. This ability depends upon the capacity to imitate and conform, which, in turn, leads to strong intergroup differences, which, in turn, create variation that allows for selection. Returning to Darwin, these abilities are among the essential ones that make a difference or, to use his own words, that make the moral sense in animals only “nearly as well developed” as ours. Moreover, it is these differences that make Darwin’s adaptive account seem at least plausible: “At all times throughout the world tribes have supplanted other tribes; and as morality is one important element in their success, the standard of morality and the number of well-endowed men will thus everywhere tend to rise and increase.” That said, Darwin’s answer is only partial here. It only provides an explanation of how there can be moral evolution, and why it might be favored. The other part of the answer comes from what it does for both the individual and the group. At the individual level, a moral system not only curbs selfish action by means of setting guidelines for action but also sets sanctions against those who violate them. At this level, targeted at the descriptive principles associated with social norms, there are similarities with animals, especially in terms of their social organizations and the constraints they impose on resource distribution: fidelity to mates, property rights, sharing, and investment in relationships. Where humans took an important turn is in the conversion between descriptive and prescriptive principles for handling distributive justice.

I stated earlier that there is a strong and weak analogy to language. The strong analogy holds that the architecture of the language faculty is structurally and functionally like morality, even though each solves a different adaptive problem. The weak analogy holds that by following the line of questioning in linguistics, we will make great strides in understanding morality. I hope that no one seriously rejects the weak analogy. Though we are certainly not ready to accept or reject the strong analogy, let me wrap up this section by taking stock. Consider, by way of analogy, a recent list of the basic facts of language described by linguists Hornstein, Nunes, and Grohmann in their book *Understanding Minimalism*:

1. Sentences are basic linguistic units.
2. Sentences are pairings of form (sound/signs) and meaning.
3. Sentences are composed of smaller expressions (words and morphemes).
4. These smaller units are composed into units with hierarchical structure, that is, phrases larger than words and smaller than sentences.

5. Sentences show displacement properties in the sense that expressions that appear in one position can be interpreted in another.
6. Language is recursive, that is, there is no upper bound on the length of sentences in any given natural language.

Given the discussion in this essay, and the work carried out in moral psychology thus far, I think we are in a reasonable position to substitute in the relevant changes for morality:

1. Events are basic moral units.
2. Events are pairings of form (actions) and meaning.
3. Events are composed of smaller expressions (intentional acts, goals, subgoals).
4. These smaller units are composed into units with hierarchical structure, that is, goals larger than subgoals, which are larger than actions.
5. Events show displacement properties in the sense that meaningful actions that appear in one position can be interpreted in another.
6. Morality is recursive, that is, there's no upper bound on the duration or number actions of an event in any given natural moral system.

I am skipping many of the critical details of this analogy, but I certainly do not think the transition seems forced. Consider it an interim report.

For those who are unimpressed by the arguments and evidence for a moral faculty with Rawlsian design specs, I leave you with one final connection to language, especially its history of theoretical upheavals and changes. The signature of progress in any science is an increasingly rich set of explanatory principles to account for the phenomenon at hand, as well as the delivery of new questions that could never have been contemplated in the past. Linguistics, as a discipline, has gone through numerous shifts in recent history, from sitting within the humanities and social sciences to stretching sideways to the natural sciences, and especially the cognitive and neurosciences. As a result of this shift, it has also witnessed a change in its approach, from seeing language as a cultural object to seeing it as a natural object, one that is as much a part of biological inquiry as the heart or eye. Within this shift, there have also been numerous changes in its theoretical constructs. These changes largely follow a similar path: an attempt to deliver the most economical, simple, and beautiful explanation of the mature state of linguistic knowledge and its acquisition.

If linguistics is any guide, and if history provides insights for what is in store, then by raising new questions about our moral faculty as I have

done here, we are on the verge of a Renaissance in our understanding of the moral domain.

2.4. *The Is of Our Ought*

As I write the last few words of this essay, the text of the *New York Times* and the voice of National Public Radio remind me of some of the moral atrocities that blanket our globe: brutal warfare in Iraq, Afghanistan, Georgia, Palestine, Israel, Sudan, and the Congo; senseless starvation and suffering in Darfur; and the loss of homes and mortgages by hardworking Americans due to the greed of fat-cat Wall Streeters. Thinking about such problems not only invites a rich sense of déjà vu but a concern that, whatever our biology is doing, it is not enough to capture our sense of what we ought to do. The *is* simply lacks the oomph to guide our reflective *ought*, or, if it has the oomph, it lacks the sophistication to get things right in a rapidly changing world that is foreign to our evolved moral sense.

As Rawls neared the end of his life, he articulated a contemplative ideal, one that captures the notion of a moralspace: “Our social world might have been different and there is hope for those at another time and place.” The operative phrase here is “might have been different.” This speaks to the idea of a set of possible moral systems, and, symmetrically, points to systems that have failed and are thus, at some level, impossible. One implication of this perspective, perhaps a tad utopian, is that our biology provides options for building possible moral systems that are sufficiently rich to allow for different attitudes, but sufficiently constrained as to block off destructive ones. In closing, I want to point to some potentially productive avenues for future exploration, a space where work on our descriptive ethics can productively contribute to progress in prescriptive ethics. In a nutshell, I want to argue what I hope is a noncontroversial point: if we are to advance a prescriptive ethics that safeguards the basic needs of the individual while allowing a plurality of ideals within and between societies, our only hope is to gain a deep understanding of human nature, characterizing how it buckles under some conditions and how it rises to virtuous achievements under others.

In Lecture I of this essay, I provided a theoretical framework for thinking about humaniqueness as well as the core principles that underpin our capacity for massive cultural expression. We can put these two issues together to think about our moral psychology, and, in particular, the debate concerning moral relativism. As I understand it, there are at least three different concerns having to do with relativism. The first is a descriptive

claim, specifically, the extent to which there is cross-cultural variability in moral attitudes and behavior. The second is a normative or prescriptive claim, specifically, whether one culture should be allowed to impose its moral norms on another culture. Those who support a relativist perspective in a normative sense think that each culture should be free to set its own moral agenda. Last, there is the metaethical concern, specifically, a question of whether there are moral truths and, if so, whether they cut across cultures (universality) or rely on the particular details of each culture. The work that my students and I have carried out is most directly concerned with the descriptive claim, but some of the work may well have relevance to the normative and metaethical claims. Let me explain.

Our work on the Internet with the MST, together with our field studies of the Mayans, licenses two conclusions: first, in the face of quite substantial variation in cultural background (education, religion), some psychological distinctions (means versus side effects) carry through and guide moral judgments; second, other distinctions are open to cross-cultural variation (actions versus omissions). This work, we believe, is an advance over previous empirical work that has tended to focus on moral behavior as opposed to judgment, and on moral issues that are coarse- as opposed to fine-grained in terms of the underlying psychological mechanisms. Thus, there are numerous anthropological studies showcasing cultural variation in the ethics surrounding sexual behavior (homosexuality, incest, infidelity), parenting (infanticide), and violence (punishment for murder, theft). But this work, though of great interest, does not allow a sufficiently fine-grained articulation of the psychological mechanisms that generate the kinds of subtle differences in moral judgment uncovered in the philosophical literature. Thus, the work we have begun—and it is only a beginning—is an attempt to run cross-cultural work with this kind of descriptive potential.

One way in which the work on descriptive ethics may prove relevant to prescriptive concerns about relativism is by addressing the argument from disagreement.¹⁹ Several authors have argued that the best explanation in support of descriptive moral relativism is metaethical moral relativism, and, in particular, the idea that there are no moral truths but rather moral rules that each culture decides. The classic counterargument to the alignment of descriptive and metaethical moral relativism is to show that

19. I owe considerable thanks to Ben Fraser for clarifying these issues to me. See J. M. Dorris and A. Plakias, "How to Argue about Disagreement," in vol. 2 of *Moral Psychology*, edited by W. Sinnott-Armstrong (Cambridge: MIT Press, 2008).

underlying the observed differences between cultures are differences in beliefs, modes of reasoning, and prejudices. To defeat this view, it is necessary to show that these differences stem from nonmoral concerns. Thus, for example, suppose that culture A believes that incest is best because of the folk view that intercourse among kin produces the fittest children. Culture B, in contrast, has a different view, based on scientific evidence that intercourse among kin creates unfit children, often characterized by neurological abnormalities. Both sets of beliefs are nonmoral, though they fuel moral judgments. As a result, it is possible to defeat the argument from disagreement by showing that, at their root, the differences rely on nonmoral concerns.

Now consider the Mayan data I presented. Like our Internet sample, the Mayans perceive a moral difference between means-based harms and foreseen side effects, with the former judged more harshly than the latter. In contrast, though our Internet sample perceives a difference between actions and omissions, the Mayans do not. Thus, the contrast between our Internet sample and the Mayans represents a case of moral disagreement that would appear indefeasible on a straightforward appeal to defusing explanations. For example, though actions are more transparently linked to causal attributions than omissions, this is a nonmoral distinction that the Mayans perceive as clearly as do the subjects on the Internet. Similarly, the Mayan failure to judge actions as worse than omissions cannot be due to problems of reasoning about scales, as they perform as predicted on the means–side effect cases, and it is also not due to educational background, as even young American children, in the four- to six-year range, perceive a difference between actions and omissions. What we suggest may drive the cross-cultural difference is the fact that this Mayan population lives in a small-scale society, whereas subjects on the Internet do not. Although the size of a culture’s population is not, in and of itself, a moral distinction, it is one that maps onto the knowledge that each individual has of the other, and, ultimately, of their responsibility for their actions. I believe this breakdown of the society’s scale is, at its core, part of the morally relevant variation. What the Mayan data imply, therefore, is that humans are endowed with an act-omit parameter. In our ancestral past, we lived in small-scale societies. This parameter thus defaulted to “off.” As societies grew larger, the parameter was turned off, as it was simply impossible to hold others responsible for either their actions or their inactions. Though we cannot be certain if this is the right analysis, it provides an enticing way of thinking about both psychological mechanisms and their evolution.

Needless to say, the Mayan data do not constitute knockdown evidence against the argument from disagreement, but they are in line with the kind of evidence one needs. More generally, these are the kinds of studies that may help link descriptive and prescriptive ethics and, ultimately, help explore the validity of moral absolutism—the possibility that there are absolute moral truths regardless of culture.

The second line of research linking descriptive with prescriptive issues comes from studies of psychopaths, as well as other clinical populations. At its core, the law has one goal: to determine whether the means by which a particular consequence was achieved deserves punishment of some kind. The law attempts to work out the means because bad consequences may arise by accident or as malicious intent; typically, only the latter invites a punishing sentence, whereas the former does not, unless there is evidence of negligence. In exploring the nature of the psychopath's mind and the link between thought and violent behavior, I raised a question concerning the cause of his actions: specifically, is the lack of morally appropriate behavior due to lack of morally appropriate emotions, lack of moral knowledge, lack of general self-control, or some combination of all three of these factors? Though the clinical diagnosis is clear, pointing to deficits in emotional processing, especially the social emotions of empathy, guilt, and remorse, as well as in self-control, tests of moral knowledge reveal few differences with healthy subjects. In particular, when psychopaths respond to moral dilemmas, cases where there are no rules to guide judgment, they not only show sensitivity to impersonal and personal cases as do healthy subjects—personal dilemmas are more often perceived as forbidden transgressions—but show no difference in their judgments with such healthy subjects. These results license the conclusion that psychopaths know right from wrong but do not care. Viewed from the *evodevo* perspective developed in Lecture I, whatever systems of the mind are responsible for building moral knowledge and guiding judgment, there is a disconnect with the systems that are responsible for motivating morally relevant behavior.

This diagnosis of psychopaths raises new issues for both law and treatment, and, in particular, the intersection between these two threads. When the law attacks a criminal case, it seeks information about the criminal's mental state, and, in particular, about whether the criminal acted knowingly or purposefully. Though one could spend an inordinate amount of time debating the meaning of these terms, the evidence presented suggests that at some level, psychopaths have a well-articulated understanding of

moral rights and wrongs, making subtle distinctions among cases based on deontological and utilitarian concerns. As such, it would be difficult to defend the claim that psychopaths kill or engage in financially corrosive relationships without knowing that what they are doing is wrong. What about the purpose behind their actions? Do they kill or extort purposefully? Answering this question requires a consideration of what psychopaths know, what they feel, and the extent to which they are capable of self-control. Given the work thus far, this trio of capacities presents a puzzle: psychopaths know right from wrong, do not feel bad when they harm others, and exhibit weak inhibitory control. On the one hand, therefore, their decision to bring about a prohibited consequence (murder, financial ruin) is intended. But given the fact that they fail to experience our species-typical emotions when considering a prohibited goal or outcome, their goal-directed behavior is significantly compromised. Added on to this problem is the fact that psychopaths show a compromised capacity for inhibitory control, acting impulsively. This combination of deficits effectively challenges the notion of purpose, suggesting instead that psychopaths lack goals, killing and extorting without purpose. That is, there is a disconnect between three systems engaged with our moral psychology: (1) a system of moral knowledge or competence that underpins our moral judgments; (2) a system of emotional experience that gives us a high when we act virtuously and a low when we lie, steal, or harm; and (3) a system that regulates our actions, either facilitating or inhibiting. Recognizing this triumvirate of processes should force a modification of the Model Penal Code, and especially its rather straightforward reliance on classic definitions of terms such as “purposefully” and “knowingly.” Similarly, given the breakdown between these processes, clinicians and neuroscientists must begin to explore how the different modules of the brain engage at both the neuronal and the neurochemical levels, with the hope that, someday soon, we will understand these interfaces and how they can be repaired.

The work presented here also has implications for education, and how to think about the developing child’s acquisition of a moral code. In many circles, education takes on a gas stationesque metaphor, each child arriving in class with an empty reservoir, ready to be filled up by the sagacious teacher. Children are, however, anything but empty tanks. They arrive with preexisting conceptions of the way the world works, and such prior knowledge constrains future acquisition. It can also facilitate future understanding, as recently demonstrated in the domain of mathematics.

Research by Halberda and colleagues indicates that individual differences in the evolutionarily ancient system of approximate magnitude estimation is correlated with performance on subsequent standardized math tests. Specifically, young children with greater acuity of discrimination for approximate calculations subsequently perform better on standardized math tests. How might similar precultural capacities operate in the moral domain? I can imagine two directions. First, parents and educators need to appreciate that children start life with the necessary building blocks to distinguish intended from accidental actions, as well as fair from unfair distributions. These building blocks constrain their interpretations of actions and events. As in studies of mathematics, educational sensitivity to such early abilities, including individual differences, might combine productively with more formal education to enhance the child's path to moral growth and development. Second, there are significant individual differences in self-control, as demonstrated by Mischel's gorgeous studies of delayed gratification. Specifically, some children are remarkably impatient in the context of waiting for delayed rewards, whereas others show self-control, bypassing the opportunity to take what is easy and immediate. These differences are predictive of future delinquency, including juvenile violence and gambling, as well as the capacity to maintain a stable marital relationship. Combine the child-as-intuitive-jurist with the variability in self-control, and we have an individual who imposes constraints on the role of experience in generating a culturally specific moral signature. This point is especially important in light of recent evidence of cross-cultural differences in performance on bargaining games among individuals living in small-scale societies, and the observation that children begin life selfish, immune to inequities. If the default human starting state is, as in chimpanzees, selfish, then, as Dawkins noted more than thirty years ago, we must teach kindness. And, if we do, it must be in the context of a biology that has the potential for a broad moralspace, capable of democratic distribution of basic needs, on the one hand, and of oligarchic and egotistical control of resources, on the other. We alone have the freedom to decide how to make use of this space. It is one of the benefits of humaniqueness, of being something other than an animal.