3 Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions

Marc D. Hauser, Liane Young, and Fiery Cushman

The thesis we develop in this essay is that all humans are endowed with a *moral faculty*. The moral faculty enables us to produce moral judgments on the basis of the causes and consequences of actions. As an empirical research program, we follow the framework of modern linguistics.[1] The spirit of the argument dates back at least to the economist Adam Smith (1759/1976) who argued for something akin to a moral grammar, and more recently, to the political philosopher John Rawls (1971). The logic of the argument, however, comes from Noam Chomsky's thinking on language specifically and the nature of knowledge more generally (Chomsky, 1986, 1988, 2000; Saporta, 1978).

If the nature of moral knowledge is comparable in some way to the nature of linguistic knowledge, as defended recently by Harman (1977), Dwyer (1999, 2004), and Mikhail (2000; in press), then what should we expect to find when we look at the anatomy of our moral faculty? Is there a grammar, and if so, how can the moral grammarian uncover its structure? Are we aware of our moral grammar, its method of operation, and its moment-to-moment functioning in our judgments? Is there a universal moral grammar that allows each child to build a particular moral grammar? Once acquired, are different moral grammars mutually incomprehensible in the same way that a native Chinese speaker finds a native Italian speaker incomprehensible? How does the child acquire a particular moral grammar, especially if her experiences are impoverished relative to the moral judgments she makes? Are there certain forms of brain damage that disrupt moral competence but leave other forms of reasoning intact? And how did this machinery evolve, and for what particular adaptive function? We will have more to say about many of these questions later on, and Hauser (2006) develops others. However, in order to flesh out the key ideas and particular empirical research paths, let us turn to some of the central questions in the study of our language faculty.

Chomsky, the Language Faculty, and the Nature of Knowing

Human beings are endowed with a language faculty—a mental "organ" that learns, perceives, and produces language. In the broadest sense, the language faculty can be thought of as an instinct to acquire a natural language (Pinker, 1994). More narrowly, it can be thought of as the set of principles for growing a language.

Prior to the revolution in linguistics ignited by Chomsky, it was widely held that language could be understood as a cultural construction learned through simple stimulus–response mechanisms. It was presumed that the human brain was more or less a blank slate upon which anything could be imprinted, including language. Chomsky, among others, challenged this idea with persuasive arguments that human knowledge of language must be guided in part by an innate faculty of the mind—the faculty of language. It is precisely because of the structure of this faculty that children can acquire language in the absence of tutelage, and even in the presence of negative or impoverished input.

When linguists refer to these principles as the speaker's "grammar," they mean the rules or operations that allow any normally developing human to unconsciously generate and comprehend a limitless range of well-formed sentences in their native language. When linguists refer to "universal grammar" they are referring to a theory about the set of all principles available to each child for acquiring a natural language. Before the child is born, she doesn't know which language she will meet, and she may even meet two if she is born in a bilingual family. However, she doesn't need to know. What she has is a set of principles and parameters that prepares her to construct different grammars that characterize the world's languages—dead ones, living ones, and those not yet conceived. The environment feeds her the particular sound patterns (or signs for those who are deaf) of the native language, thereby turning on the specific parameters that characterize the native language.

From these general problems, Chomsky and other generative grammarians suggested that we need an explicit characterization of the language faculty, what it is, how it develops within each individual, and how it evolved in our species, perhaps uniquely (Anderson & Lightfoot, 2000; Fitch, Hauser, & Chomsky, 2005; Hauser, Chomsky, & Fitch, 2002; Jackendoff, 2002; Pinker, 1994). We discuss each of these issues in turn.

What Is It?

The faculty of language is designed to handle knowledge of language. For English speakers, for instance, the faculty of language provides the princi-

ples upon which our knowledge of the English language is constructed. To properly understand what it means to know a language, we must distinguish between *expressed* and *operative* knowledge. Expressed knowledge includes what we can articulate, including such things as our knowledge that a fly ball travels a parabolic arc describable by a quadratic mathematical expression. Operative knowledge includes such things as our knowledge of how to run to just the right spot on a baseball field in order to catch a fly ball. Notice that in the case of baseball, even though our expressed knowledge about the ball's parabolic trajectory might be used to inform us about where to run if we had a great deal of time and sophisticated measuring instruments, it is of little use in the practical circumstances of a baseball game. In order to perform in the real world, our operative knowledge of how to run to the right spot is much more useful. Our brain must be carrying out these computations in order for us to get to the right spot even though, by definition, we can't articulate the principles underlying this knowledge. In the real-world case of catching a baseball, we rely on operative as opposed to expressed knowledge.

One of the principle insights of modern linguistics is that knowledge of language is operative but not expressed. When Chomsky generated the sentence "Colorless green ideas sleep furiously," he intentionally produced a string of words that no one had ever produced before. He also produced a perfectly grammatical and yet meaningless sentence. Most of us don't know what makes Chomsky's sentence, or any other sentence, grammatical. We may express some principle or rule that we learned in grammar school, but such expressed rules are rarely sufficient to explain the principles that actually underlie our judgments. It is these unconscious or operative principles that linguists discover—and that never appear in the schoolmarm's textbook—that account for the patterns of linguistic variation and similarities. For example, every speaker of English knows that "Romeo loves Juliet" is a well-formed sentence, while "Him loves her" is not. Few speakers of English know why. Few native speakers of English would ever produce this last sentence, and this includes young toddlers just learning to speak English. When it comes to language, therefore, what we think we know pales in relation to what our minds actually know. Similarly, unconscious principles underlie certain aspects of mathematics, music, object perception (Dehaene, 1997; Jackendoff, 2005; Lerdahl & Jackendoff, 1996; Spelke, 1994), and, we suggest, morality (Hauser, 2006; Mikhail, 2000, in press).

Characterizing our knowledge of language in the abstract begins to answer the question "What is the faculty of language," but in order to achieve a more complete answer we want to explain the kinds of processes

of the mind/brain that are specific to language as opposed to shared with other problem-oriented tasks including navigation, social relationships, object recognition, and sound localization. The faculty of language's relationship to other mind-internal systems can be described along two orthogonal dimensions: whether the mechanism is necessary for language and whether the mechanism is unique to language. For example, we use our ears when we listen to a person speaking and when we localize an ambulance's siren, and deaf perceivers of sign language accomplish linguistic understanding without using their ears at all. Ears, therefore, are neither necessary for nor unique to language. However, once sound passes from our ears to the part of the brain involved in decoding what the sound is and what to do with it, separate cognitive mechanisms come in to play, one for handling speech, the other nonspeech. Speech-specific perceptual mechanisms are unique to language but still not necessary (again, consider the deaf).

Once the system detects that we are in a language mode, either producing utterances or listening to them, a system of rules is engaged, organizing meaningless sound and/or gesture sequences (phonemes) into meaningful words, phrases, and sentences, and enabling conversation either as internal monologue or external dialogue. This stage of cognitive processing is common to both spoken and sign language. The hierarchical structure of language, together with its recursive and combinatorial operations, as well as interfaces to phonology and semantics, appear to be unique properties of language *and* necessary for language. We can see, then, that the faculty of language is comprised of several different types of cognitive mechanisms: those that are unique versus those that are shared and those that are necessary versus those that are optionally recruited.

To summarize, we have now sketched the abstract system of knowledge that characterizes the faculty of language, and we have also said something about the different ways in which cognitive mechanisms can be integrated into the faculty of language. There remains one more important distinction that will help us unpack the question "What is the faculty of language": the distinction between linguistic competence, or what the language faculty enables, and linguistic performance, or what the rest of the brain and the environment constrain. Language competence refers to the unconscious and inaccessible principles that make sentence production and comprehension possible. What we say, to whom, and how is the province of linguistic performance and includes many other players of the brain, and many factors external to the brain, including other people,

institutions, weather, and distance to one's target audience. When we speak about the language faculty, therefore, we are speaking about the normal, mature individual's *competence* with the principles that underlie her native language. What this individual chooses to say is a matter of her *performance* that will be influenced by whether she is tired, happy, in a fight with her lover, or addressing a stadium-filled audience.

How Does It Develop?

To answer this question, we want to explain the child's path to a mature state of language competence, a state that includes the capacity to create a limitless range of meaningful sentences and understand an equally limitless range of sentences generated by other speakers of the same language. Like all biological phenomena, the development of language is a complex interaction between innate structure, maturational factors, and environmental input. While it is obvious that much of language is learned—for instance, the arbitrary mapping between sound and concept—what is less obvious is that the learning of language is only possible if the learner is permitted to make certain initial assumptions. This boils down to a question of the child's initial state—of her unconscious knowledge of linguistic principles prior to exposure to a spoken or signed language. It has to be the case that some innate structure is in place to guide the growth of a particular language, as no other species does the same (even though cats and dogs are exposed to the same stuff), and the input into the child is both impoverished and replete with ungrammatical structure that the child never repeats.

Consider the observation that in spoken English, people can use two different forms of the verb "is" as in "Frank is foolish" and "Frank's foolish." We can't, however, use the contracted form of *is* wherever we please. For example, although we can say "Frank is more foolish than Joe is," we can't say "Frank is more foolish than Joe's." How do we know this? No one taught us this rule. No one listed the exceptions. Nonetheless, young children never use the contracted form in an inappropriate place. The explanation, based on considerable work in linguistics (Anderson & Lightfoot, 2000), is that the child's initial state includes a principle for verb contraction—a rule that says something like "*'s* is too small a unit of sound to be alone; whenever you use the contracted form, follow it up with another word." The environment—the sound pattern of English—triggers the principle, pulling it out of a hat of principles as if by magic. The child is born knowing the principle, even though she is not consciously aware of the knowledge she holds. The principle is operative but not expressed.

There are two critical points to make about the interplay between language and the innate principles and parameters of language learners. First, the principles and parameters are what make language learning possible. By guiding children's expectations about language in a particular fashion, the principles and parameters allow children to infer a regular system with infinite generative capacity from sparse, inconsistent, and imperfect evidence. However, the principles and parameters do not come for free, and this brings us to the second point: the reason that principles and parameters make the child's job of learning easier is because they restrict the range of possible languages. In the example described above, the price of constraining a child's innate expectations about verb contraction is that it is impossible for any language to violate that expectation.

To summarize, the development of the language faculty is a complex interaction of innate and learned elements. Some elements of our knowledge of language are precisely specified principles, invariant between languages. Other elements of our knowledge of language are parametrically constrained to a limited set of options, varying within this set from language to language. Finally, some elements of our knowledge of language are unconstrained and vary completely from language to language. We note here that, although we have leaned on the principles and parameters view of language, this aspect of our argument is not critical to the development of the analogy between language and morality. Other versions of the generative grammar perspective would be equally appropriate, as they generally appeal to language-specific, universal computations that constrain the range of cultural variation and facilitate acquisition.

How Did It Evolve?

To answer this question, we look to our history. Which components of our language faculty are shared with other species, and which are unique? What problems did our ancestors face that might have selected for the design features of our language faculty? Consider the human child's capacity to learn words. Much of word learning involves vocal imitation. The child hears her mother say, "Do you want candy?" and the child says "Candy." "Candy" isn't encoded in the mind as a string of DNA. But the capacity to imitate sounds is one of the human child's innate gifts. Imitation is not specific to the language faculty, but without it, no child could acquire the words of his or her native language, reaching a stunning level of about 50,000 for the average high school graduate. To explore whether vocal imitation is unique to humans, we look to other species. Although we share 98% of our genes in common with chimpanzees, chimpanzees

show no evidence of vocal imitation. The same goes for all of the other apes and all of the monkeys. What this pattern tells us is that humans evolved the capacity for vocal imitation some time after we broke off from our common ancestor with chimpanzees—something like 6 to 7 million years ago. However, this is not the end of our exploration. It turns out that other species, more distantly related to us than any of the nonhuman primates, are capable of vocal imitation: all Passerine songbirds, parrots, hummingbirds, dolphins, and some whales. What this distribution tells us is that vocal imitation is not unique to humans. It also tells us, again, that vocal imitation in humans didn't evolve from the nonhuman primates. Rather, vocal imitation evolved independently in humans, some birds, and some marine mammals.

To provide a complete description of the language faculty, addressing each of the three questions discussed, requires different kinds of evidence. For example, linguists reveal the deep structure underlying sentence construction by using grammaticality judgments and by comparing different languages to reveal commonalities that cut across the obvious differences. Developmental psychologists chart the child's patterns of language acquisition, exploring whether the relevant linguistic input is sufficient to account for their output. Neuropsychologists look to patients with selective damage, using cases where particular aspects of language are damaged while others are spared, or where language remains intact and many other cognitive faculties are impaired. Cognitive neuroscientists use imaging techniques to understand which regions of the brain are recruited during language processing, attempting to characterize the circuitry of the language organ. Evolutionary biologists explore which aspects of the language faculty are shared with other species, attempting to pinpoint which components might account for the vast difference in expressive power between our system of communication and theirs. Mathematical biologists use models to explore how different learning mechanisms might account for patterns of language acquisition, or to understand the limiting conditions for the evolution of a universal grammar. This intellectual collaboration is beginning to unveil what it means to know a particular language and to use it in the service of interacting with the world. Our goal is to sketch how similar moves can be made with respect to our moral knowledge.

Rawls and the Linguistic Analogy

In 1950, Rawls completed his PhD, focusing on methodological issues associated with ethical knowledge and with the characterization of a person's

moral worth. His interest in our moral psychology continued up until the mid-1970s, focusing on the problem of justice as fairness, and ending quite soon after the publication of *A Theory of Justice*.

Rawls was interested in the idea that the principles underlying our intuitions about morality may well be unconscious and inaccessible.[2] This perspective was intended to parallel Chomsky's thinking in linguistics. Unfortunately, those writing about morality in neighboring disciplines, especially within the sciences, held a different perspective. The then dominant position in developmental psychology, championed by Piaget and Kohlberg, was that the child's moral behavior is best understood in terms of the child's articulations of moral principles. Analogizing to language, this would be equivalent to claiming that the best way to understand a child's use of verb contraction is to ask the child why you can say "Frank is there" but can't ask "Where Frank's?", presuming that the pattern of behavior must be the consequence of an articulatable rule.

The essence of the approach to morality conceived by Piaget, and developed further by Kohlberg, is summarized by a simple model: the perception of an event is followed by reasoning, resulting finally in a judgment (see figure 3.1); emotion may emerge from the judgment but is not causally related to it. Here, actions are evaluated by reflecting upon specific principles and using this reflective process to rationally deduce a specific judgment. When we deliver a moral verdict, it is because we have considered different possible reasons for and against a particular action, and based on this deliberation, alight upon a particular decision. This model might be termed "Kantian," for although Kant never denied the role of intuition in our moral psychology, he is the moral philosopher who carried the most weight with respect to the role of rational deliberation about what one ought to do.

The Piaget/Kohlberg tradition has provided rich and reliable data on the moral stages through which children pass, using their justifications as

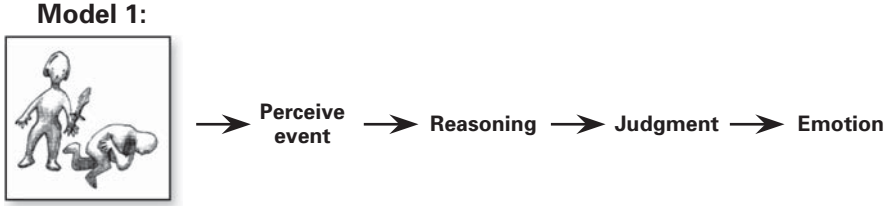

Figure 3.1
The Kantian creature and the deliberate reasoning model

primary evidence for developmental change. In recent years, however, a number of cognitive and social psychologists have criticized this perspective (Macnamara, 1990), especially its insistence that the essence of moral psychology is *justification* rather than *judgment*. It has been observed that even fully mature adults are sometimes unable to provide any sufficient justification for strongly felt moral intuitions, a phenomenon termed "moral dumbfounding" (Haidt, 2001). This has led to the introduction of a second model, characterized most recently by Haidt (2001) as well as several other social psychologists and anthropologists (see figure 3.2). Here, following the perception of an action or event, there is an unconscious emotional response which immediately causes a moral judgment; reasoning is an afterthought, offering a post hoc rationalization of an intuitively generated response. We see someone standing over a dead person and we classify this as murder, a claim that derives from a pairing between any given action and a classification of morally right or wrong. Emotion triggers the judgment. We might term this model "Humean," after the philosopher who famously declared that reason is "slave to the passions"; Haidt calls it the social intuitionist model.

A second recent challenge to the Piaget/Kohlberg tradition is a hybrid between the Humean and Kantian creatures, a blend of unconscious emotions and some form of principled and deliberate reasoning (see figure 3.3); this view has most recently been championed by Damasio based on neurologically impaired patients (S. W. Anderson, Bechara, Damasio, Tranel, & Damasio, 1999; Damasio, 1994; Tranel, Bechara, & Damasio, 2000) and by Greene (this volume) based on neuroimaging work (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).[3] These two systems may converge or diverge in their assessment of the situation, run in parallel or in sequence, but both are precursors to the judgment; if they diverge, then some other mechanism must intrude, resolve the conflict, and generate a judgment. On Damasio's view,

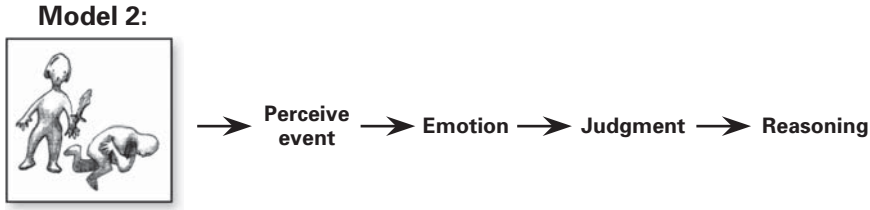

Figure 3.2
The Humean creature and the emotional model.

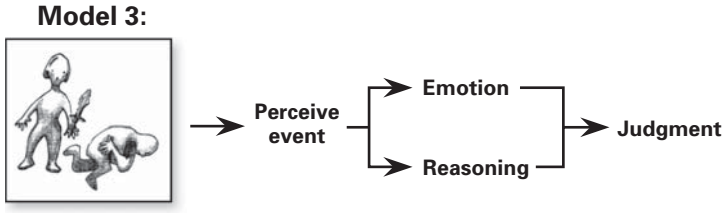

Figure 3.3
A mixture of the Kantian and Humean creatures, blending the reasoning and emotional models.

every moral judgment includes both emotion and reasoning. On Greene's view, emotions come into play in situations of a more personal nature and favor more deontological judgments, while reason comes into play in situations of a more impersonal nature and favors more utilitarian judgments.

Independent of which account turns out to be correct, this breakdown reveals a missing ingredient in almost all current theories and studies of our moral psychology. It will not do merely to assign the role of moral judgment to reason, emotion, or both. We must describe computations underlying the judgments that we produce. In contrast to the detailed work in linguistics focusing on the principles that organize phonology, semantics, and syntax, we lack a comparably detailed analysis of how humans and other organisms perceive actions and events in terms of their causal-intentional structure and the consequences the ensue for self and other. As Mikhail (2000; in press), Jackendoff (2005), and Hauser (2006) have noted, however, actions represent the right kind of unit for moral appraisal: discrete and combinable to create a limitless range of meaningful variation.

To fill in this missing gap, we must characterize knowledge of moral codes in a manner directly comparable to the linguist's characterization of knowledge of language. This insight is at the heart of Rawls's linguistic analogy. Rawls (1971) writes, "A conception of justice characterizes our moral sensibility when the everyday judgments we make are in accordance with its principles" (p. 46). He went on to sketch the connection to language:

> A useful comparison here is with the problem of describing the sense of grammaticalness that we have for the sentences of our native language. In this case, the aim is to characterize the ability to recognize well-formed sentences by formulating clearly expressed principles which make the same discriminations as the native speaker. This is a difficult undertaking which, although still unfinished, is known

to require theoretical constructions that far outrun the ad hoc precepts of our explicit grammatical knowledge. A similar situation presumably holds in moral philosophy. There is no reason to assume that our sense of justice can be adequately characterized by familiar common sense precepts, or derived from the more obvious learning principles. A correct account of moral capacities will certainly involve principles and theoretical constructions which go beyond the norms and standards cited in every day life. (pp. 46–47)

We are now ready, at last, to appreciate and develop Rawls's insights, especially his linguistic analogy. We are ready to introduce a "Rawlsian creature," equipped with the machinery to deliver moral verdicts based on principles that may be inaccessible (see figure 3.4; Hauser, 2006); in fact, if the analogy to language holds, the principles will be operative but not expressed, and only discoverable with the tools of science. There are two ways to view the Rawlsian creature in relationship to the other models. Minimally, each of the other models must recognize an appraisal system that computes the causal-intentional structure of an agent's actions and the consequences that follow. More strongly, the Rawlsian creature provides the sole basis for our judgments of morally forbidden, permissible, or obligatory actions, with emotions and reasoning following. To be clear: the Rawlsian model does not deny the role of emotion or reasoning. Rather, it stipulates that any process giving rise to moral judgments must minimally do so on the basis of some system of analysis and that this analysis constitutes the heart of the moral faculty. On the stronger view, the operative principles of the moral faculty do all the heavy lifting, generating a moral verdict that may or may not generate an emotion or a process of rational and principled deliberation.

One way to develop the linguistic analogy is to raise the same questions about the moral faculty that Chomsky and other generative grammarians raised for the language faculty. With the Rawlsian creature in mind, let us unpack the ideas.

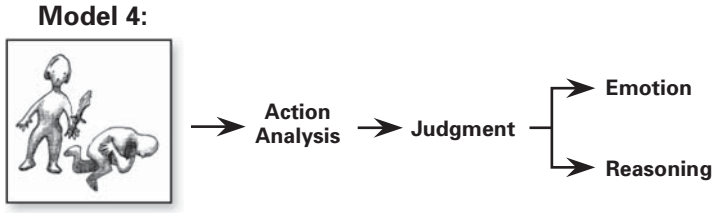

Figure 3.4
The Rawlsian creature and action analysis model.

What Is It?

Rawls argued that because our moral faculty is analogous to our linguistic faculty, we can study it in some of the same ways. In parallel with the linguist's use of grammaticality judgments to uncover some of the principles of language competence, students of moral behavior might use *morality* judgments to uncover some of the principles underlying our judgments of what is morally right and wrong.[4] These principles might constitute the Rawlsian creature's universal moral grammar, with each culture expressing a specific moral grammar. As is the case for language, this view does not deny cultural variation. Rather, it predicts variation based on how each culture switches on or off particular parameters. An individual's moral grammar enables him to unconsciously generate a limitless range of moral judgments within the native culture.

To flesh out these general comments, consider once again language. The language faculty takes as input discrete elements that can be combined and recombined to create an infinite variety of meaningful expressions: phonemes ("distinctive features" in the lingo of linguistics) for individuals who can hear, signs for those who are deaf. When a phoneme is combined with another, it creates a syllable. When syllables are combined, they can create words. When words are combined, they can create phrases. And when phrases are combined, they can create sentences that form the power of *The Iliad*, *The Origin of Species,* or *Mad Magazine*. Actions appear to live in a parallel hierarchical universe. Like phonemes, many actions may lack meaning depending upon context: lifting your elbow off the table, raising your ring finger, flexing your knee. Actions, when combined, are often meaningful: lifting your elbow and swinging it intentionally into someone's face, raising your ring finger to receive a wedding band, flexing your knee in a dance. Like phonemes, when actions are combined, they do not blend; individual actions maintain their integrity. When actions are combined, they can represent an agent's goals, his means, and the consequences of his action and inaction. When a series of subgoals are combined, they can create events, including the *Nutcracker* Ballet, the World Series, or the American Civil War. Because actions and events can be combined into an infinite variety of strings, it would be a burdensome and incomplete moral theory that attempted to link a particular judgment with each particular string individually. Instead of recalling that it was impermissible for John to attack Fred and cause him pain, we recall a principle with abstract placeholders or variables such as AGENT, INTENTION, BELIEF, ACTION, RECEIVER, CONSEQUENCE, MORAL EVALUATION. For example, the principle might generate the evaluation "Impermissible"

when intention is extended over an action that is extended over a harm (see figure 3.5). In reality, the principle will be far more complicated and abstract and include other parameters. See Mikhail (2000; in press) for one version of how such representational structures might be constructed and evaluated in more detail.

By breaking down the principle into components, we achieve a second parallel with language: to attain its limitless range of expressive power, the moral faculty must take a finite set of elements and recombine them into new, meaningful expressions or principles. These elements must not blend like paint. Combining red and white paint yields pink. Although this kind of combination gives paint, and color more generally, a vast play space for variation, once combined we can no longer recover the elements. Each contributing element or primary color has lost its individually distinctive contribution. Not so for language or morality. The words in "John kisses Mary" can be recombined to create the new sentence "Mary kisses John." These sentences have the same elements (words), and their ordering is uniquely responsible for meaning. Combining these elements does not, however, dilute or change what each means. John is still the same person in these two sentences, but in one he is the SUBJECT and in the other he is the OBJECT. The same is true of morality and our perception of the causes and consequences of actions. Consider the following two events: "Mother gratuitously hits 3-year-old son" versus "Three-year-old son gratuitously hits mother." The first almost certainly invokes a moral evaluation that harming is forbidden, while the second presumably doesn't. In the first case we imagine a malignant cause, whereas in the second we

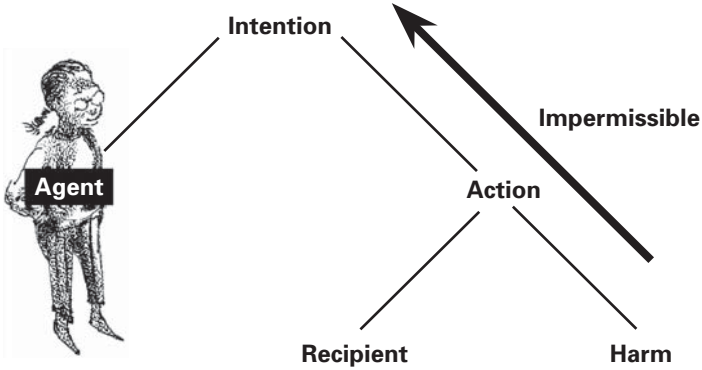

Figure 3.5
Some components of the causes and consequences of morally relevant actions.

imagine a benign cause, focused on the boy's frustration or inability to control anger.

Added on to this layer of description is another, building further on the linguistic analogy: if there is a specialized system for making moral judgments, then damage to this system should cause a selective deficit, specifically, deterioration of the moral sensibilities. To expose our moral knowledge, we must look at the nature of our action and event perception, the attribution of cause and consequence, the relationship between judgment and justification, and the extent to which the mechanisms that underlie this process are specialized for the moral faculty or shared with other systems of the mind. We must also explore the possibility that although the principles of our moral faculty may be functionally imprisoned, cloistered from the system that leads to our judgments, they may come to play a role in our judgments once uncovered. In particular, and highlighting a potentially significant difference between language and morality, once detailed analyses uncover some of the relevant principles and parameters, and make these known, we may use them in our day-to-day behavior, consciously, and based on reasoning. In contrast, knowing the abstract principles underlying certain aspects of language plays no role in what we say, and this is equally true of distinguished linguists.

Before moving further, let us make two points regarding the thesis we are defending. First, as Bloom (2004; Pizarro & Bloom, 2003) has argued and as Haidt (2001) and others have acknowledged, it would be foolish to deny that we address certain moral dilemmas by means of our conscious, deliberate, and highly principled faculty of reasoning, alighting upon a judgment in the most rational of ways. This is often what happens when we face new dilemmas that we are ill equipped to handle using intuitions. For example, most people don't have unconsciously generated intuitions, emotionally mediated or not, about stem cell research or the latest technologies for in vitro fertilization, because they lack the relevant details; some may have strong intuitions that such technologies are evil because they involve killing some bit of life or modifying it in some way, independent of whether they have knowledge of the actual techniques, including their costs and benefits. To form an opinion of these biomedical advances that goes beyond their family resemblance to other cases of biological intervention, most people want to hear about the details, understand who or what will be affected and in what ways, and then, based on such information, reason through the possibilities. Of course, once one has this information, it is then easy to bypass all the mess and simply judge such cases as permissible or forbidden. One might, for example, decide,

without reasoning, that anything smelling of biomedical engineering is just evil. The main point here is that by setting up these models, we establish a framework for exploring our moral psychology.

The second point builds on the first. On the view that we hold, simplified by model 4 and the Rawlsian creature, there are strong and weak versions. The strong version provides a direct challenge to all three alternative models by arguing that prior to any emotion or process of deliberate reasoning, there must be some kind of unconscious appraisal mechanism that provides an analysis of the causes and consequences of action. This system then either does or doesn't trigger emotions and deliberate reasoning. If it does trigger these systems, they arise downstream, as a result of the judgment. Emotion and deliberate reasoning are not causally related to our initial moral judgments but, rather, are caused by the judgment. On this view, the appraisal system represents our moral competence and is responsible for the judgment. Emotion, on the other hand, is part of our moral performance. Emotions are not specific to the moral domain, but they interface with the computations that are. On this view, if we could go into the brain and turn off the emotional circuits (as arises at some level in psychopathy as well as with patients who have incurred damage to the orbitofrontal cortex; see below), we would leave our moral competence intact (i.e., moral judgments would be normal), but this would cause serious deficits with respect to moral behavior. In contrast, for either models 1 or 3, turning off the emotional circuitry would cause serious deficits for both judgment and behavior. On the weaker version of model 4, there is minimally an appraisal system that analyzes the causes and consequences of actions, leading to an emotion or process of deliberate reasoning. As everyone would presumably acknowledge, by setting our sights on the appraisal system, we will uncover its operative principles as well as its role in the causal generation of moral judgments.

How Does the Moral Faculty Develop?

To answer this question, we need an understanding of the principles (specific grammar in light of the linguistic analogy) guiding an adult's judgments. With these principles described, we can explore how they are acquired.

Rawls, like Chomsky, suggests that we may have to invent an entirely new set of concepts and terms to describe moral principles. Our more classic formulations of universal rules may fail to capture the mind's computations in the same way that grammar school grammar fails to capture the principles that are part of our language faculty. For example, a commonsense

approach to morality might dictate that all of the following actions are forbidden: killing, causing pain, stealing, cheating, lying, breaking promises, and committing adultery. However, these kinds of moral absolutes stand little chance of capturing the cross-cultural variation in our moral judgments. Some philosophers, such as Bernard Gert (1998, 2004) point out that like other rules, moral rules have exceptions. Thus, although killing is generally forbidden in all cultures, many if not all cultures recognize conditions in which killing is permitted or at least justifiable. Some cultures even support conditions in which killing is obligatory: in several Arabic countries, if a husband finds his wife in flagrante delicto, the wife's relatives are expected to kill her, thereby erasing the family's shame. Historically, in the American South, being caught in flagrante delicto was also a mark of dishonor, but it was up to the husband to regain honor by killing his spouse. In these cultures, killing is permissible and, one might even say, obligatory. What varies cross-culturally is how the local system establishes how to right a wrong. For each case, then, we want to ask: What makes these rules universal? What aspects of each rule or principle allow for cultural variation? Are there parameters that, once set, establish the differences between cultures, constraining the problem of moral development? Do the rules actually capture the relationship between the nature of the relevant actions (e.g., HARMING, HELPING), their causes (e.g., INTENDED, ACCIDENTAL), and consequences (e.g., DIRECT, INDIRECT)? Are there hidden principles, operating unconsciously, but discoverable with the tools of science? If, as Rawls intuited, the analogy between morality and language holds, then by answering these questions we will have gained considerable ground in addressing the problems of both descriptive and explanatory adequacy.

The hypothesis here is simple: our moral faculty is equipped with a universal set of principles, with each culture setting up particular exceptions by means of tweaking the relevant parameters. We want to understand the universal aspects as well as the degree of variation, what allows for it, and how it is constrained. Many questions remain open. Does the child's environment provide her with enough information to construct a moral grammar, or does the child show competences that go beyond her exposure? For example, does the child generate judgments about fairness and harm in the absence of direct pedagogy or indirect learning by watching others? If so, then this argues in favor of an even stronger analogy to language, in which the child produces grammatically structured and correct sentences in the absence of positive evidence and despite negative evidence. Thus, from an impoverished environment, the child generates a

rich output of grammatical utterances in the case of language, and judgments about permissible actions in the case of morality. Further, in the same way that we rapidly and effortlessly acquire our native language, and then slowly and agonizingly acquire second languages later in life, does the acquisition of moral knowledge follow a similar developmental path? Do we acquire our native moral norms with ease and without instruction, while painstakingly trying to memorize all the details of a new culture's mores, recalling the faux pas and punishable violations by writing them down on index cards?

How Did the Moral Faculty Evolve?

Like language, we can address this question by breaking down the moral faculty into its component parts and then exploring which components are shared with other animals and which are unique to our own species. Although it is unlikely that we will ever be able to ask animals to make ethicality judgments, we can ask about their expectations concerning rule followers and violators, whether they are sensitive to the distinction between an intentional and an accidental action, whether they experience some of the morally relevant emotions, and, if so, how they play a role in their decisions. If an animal is incapable of making the intentional–accidental distinction, then it will treat all consequences as the same, never taking into account its origins: seeing a chimpanzee fall from a tree and injure a group member is functionally equivalent to seeing a chimpanzee leap out of a tree and injure a group member; seeing an animal reach out and hand another a piece of food is functionally the same as seeing an animal reach out for its own food and accidentally dropping a piece into another's lap. Finding parallels are as important as finding differences, as both illuminate our evolutionary path, especially what we inherited and what we invented. Critically, in attempting to unravel the architecture of the moral faculty, we must understand what is uniquely human and what is unique to morality as opposed to other domains of knowledge. A rich evolutionary approach is essential.

A different position concerning the evolution of moral behavior was ignited under the name "sociobiology" in the 1970s and still smolders in disciplines ranging from biology to psychology to economics. This position attempts to account for the adaptive value of moral behavior. Sociobiology's primary tenet was that our actions are largely selfish, a behavioral strategy handed down to us over evolution and sculpted by natural selection; the unconscious demons driving our motives were masterfully designed replicators—selfish genes. Wilson (1975, 1998) and other

sociobiologists writing about ethics argued that moral systems evolved to regulate individual temptation, with emotional responses designed to facilitate cooperation and incite aggression toward those who cheat. This is an important proposal, but it is not a substitute for the Rawlsian position. Rather, it focuses on a different level or kind of causal problem. Whereas Rawls was specifically interested in the mechanisms underlying our moral psychology (both how we act and how we think we ought to act), Wilson was interested in the adaptive significance of such psychological mechanisms. Questions about mechanism should naturally lead to questions about adaptive significance. The reverse is true as well. The important point is to keep these perspectives in their proper place, never seeing them as alternative approaches to answering a question about moral behavior, or any other kind of behavior. They are complementary approaches.

We want to stress that at some level, there is nothing at all radical about this approach to understanding our moral nature. In characterizing the moral faculty, our task is to define its anatomy, specifying what properties of the mind/brain are specific to our moral judgments and what properties fall outside its scope but nonetheless play an essential supporting role. This task is no different from that involved in anatomizing other parts of our body. When anatomists describe a part of the body, they define its location, size, components, and function. The heart is located between your lungs in the middle of your chest, behind and slightly to the left of your breastbone; it is about the size of an adult's fist, weighs between 7 and 15 ounces, and consists of four chambers with valves that operate through muscle contractions; the function of the heart is to pump blood through the circulatory system of the body. Although this neatly describes the heart, it makes little sense to discuss this organ without mentioning that it is connected to other parts of the body and depends upon our nutrition and health for its proper functioning. Furthermore, although the muscles of the heart are critical for its pumping action, there are no heart-specific muscles. Anatomizing our moral faculty provides a similar challenge. For example, we would not be able to evaluate the moral significance of an action if every event perceived or imagined flitted in and out of memory without pausing for evaluation. But based on this observation, it would be incorrect to conclude that memory is a specific component of our moral anatomy. Our memories are used for many aspects of our lives, including learning how to play tennis, recalling our first rock concert, and generating expectations about a planned vacation to the Caribbean. Some of these memories reference particular aspects of our personal lives (autobiographical information about our first dentist appointment), some allow us

to remember earlier experiences (episodic recall for the smell of our mother's apple pie), some are kept in long-term storage (e.g., travel routes home), and others are short-lived (telephone number from an operator), used only for online work. Of course memories are also used to recall our own actions that were wrong, to feel bad about them, and to assess how we might change in order to better our moral standing. Our memory systems are therefore part of the support team for moral judgments, but they are not specific to the moral faculty. The same kind of thinking has to be applied to other aspects of the mind.

This is a rough sketch of the linguistic analogy, and the core issues that we believe are at stake in taking it forward, both theoretically and empirically; for a more complete treatment, see Hauser (2006). We turn next to some of the empirical evidence, much of which is preliminary.

Uncommon Bedfellows: Intuition Meets Empirical Evidence

Consider an empirical research program based on the linguistic analogy, aimed at uncovering the descriptive principles of our moral faculty. There are at least two ways to proceed. On the one hand, it is theoretically possible that language and morality will turn out to be similar in a deep sense, and thus, many of the theoretical and methodological moves deployed for the one domain will map onto the other. For example, if our moral faculty can be characterized by a universal moral grammar, consisting of a set of innately specified and inaccessible principles for building a possible moral system, then this leads to specific experiments concerning the moral acquisition device, its relative encapsulation from other faculties, and the ways in which exposure to the relevant moral data sets particular parameters. Under this construal, we distinguish between operative and expressed principles and expect a dissociation between our competence and performance—between the knowledge that guides our judgments of right and wrong and the factors that guide what we actually say or do; when confronted with a moral dilemma, what we say about this case or what we actually would do if confronted by it in real life may or may not map on to our competence. On the other hand, the analogy to language may be weak but may nonetheless serve as an important guide to empirical research, opening doors to theoretically distinctive questions that, to date, have few answers. The linguistic analogy has the potential to open new doors because prior work in moral psychology, which has generally failed to make the competence–performance distinction (Hauser, 2006; Macnamara, 1990; Mikhail, 2000), has focused on either principled reasoning or emotion as

opposed to the causal structure of action and has yet to explore the possibility of a universal set of principles and parameters that may constrain the range of culturally possible moral systems. In this section, we begin with a review of empirical findings that, minimally, provide support for the linguistic analogy in a weak sense. We then summarize the results and lay out several important directions for future research, guided by the kinds of questions that an analogy to language offers.

Judgment, Justification, and Universality

Philosophers have often used so-called "fantasy dilemmas" to explore how different parameters push our judgments around, attempting to derive not only descriptive principles but prescriptive ones. We aim to uncover whether the intuitions guiding the professional philosopher are shared with others lacking such background and assess which features of the causal structure of action are relevant to subjects' judgments, the extent to which cultural variables impinge upon such judgments, and the degree to which people have access to the principles underlying their assessments of moral actions.

To gather observations, and take advantage of philosophical analysis, we begin with the famous trolley problem (Foot, 1967; Thomson, 1970) and its family of mutants. Our justification for using artificial dilemmas, and trolley problems in particular, is threefold. First, philosophers (Fischer & Ravizza, 1992; Kamm, 1998b) have scrutinized cases like these, thereby leading to a suite of representative parameters and principles concerning the causes and consequences of action. Second, philosophers designed these cases to mirror the general architecture of real-world ethical problems, including euthanasia and abortion. In contrast to real-world cases, where there are already well-entrenched beliefs and emotional biases, artificial cases, if well designed, preserve the essence of real-world phenomena while removing any prior beliefs or emotions. Ultimately, the goal is to use insights derived from artificial cases to inform real-world problems (Kamm, 1998b), with the admittedly difficult challenge of using descriptive generalizations to inform prescriptive recommendations.[5] Third, and paralleling work in the cognitive sciences more generally, artificial cases have the advantage that they can be systematically manipulated, presented to subjects for evaluation, and then analyzed statistically with models that can tease apart the relative significance of different parametric variations. In the case of moral dilemmas, and the framework we advocate more specifically, artificial cases afford the opportunity to manipulate details of the

dilemma. Although a small number of cognitive scientists have looked at subjects' judgments when presented with trolleyesque problems, the focus has been on questions of evolutionary significance (how does genetic relatedness influence harming one to save many?) or the relationship between emotion and cognition (Greene et al., 2001, 2004; O'Neill & Petrinovich, 1998; Petrinovich, O'Neill, & Jorgensen, 1993). In contrast, Mikhail and Hauser have advocated using these cases to look at the computational operations that drive our judgments (Hauser, 2006; Mikhail, 2000; Mikhail, in press; Mikhail, Sorrentino, & Spelke, 1998).

We have used new Web-based technologies with a carefully controlled library of moral dilemmas to probe the nature of our appraisal system; this approach has been designed to collect a large and cross-culturally diverse sample of responses. Subjects voluntarily log on to the Moral Sense Test (MST) at moral.wjh.edu, enter demographic and cultural background information, and finally turn to a series of moral dilemmas. In our first round of testing, subjects responded to four trolley problems and one control (Hauser, Cushman, Young, Jin, & Mikhail, 2006). Controls entailed cases with no moral conflict, designed to elicit predictable responses if subjects were both carefully reading the cases and attempting to give veridical responses. For example, we asked subjects about the distribution of a drug to sick patients at no cost to the hospital or doctor and with unambiguous benefits to the patients. The four trolley problems are presented below and illustrated in figure 3.6;[6] during the test, we did not give subjects these schematics, though for the third and fourth scenarios, we accompanied the text of the dilemma with much simpler drawings to facilitate comprehension. After these questions were answered, we then asked subjects to justify two cases in which they provided different moral judgments; for some subjects, this was done within a session, whereas for others, it was done across sessions separated by a few weeks. In the data presented below, we focus on subjects' responses to the first dilemma presented to them during the test; this restricted analysis is intentional, designed to eliminate the potential confounds of not only order effects but the real possibility that as subjects read and think about their answers to prior dilemmas they may well change their strategies to guarantee consistency. Though this is of interest, we put it to the side for now.

Scenario 1 Denise is a passenger on a trolley whose driver has just shouted that the trolley's brakes have failed, and who then fainted of the shock. On the track ahead are five people; the banks are so steep that they will not be able to get off the track in time. The track has a side track leading off to the right, and Denise can turn the trolley onto it. Unfortunately there is one person on the right hand track. Denise

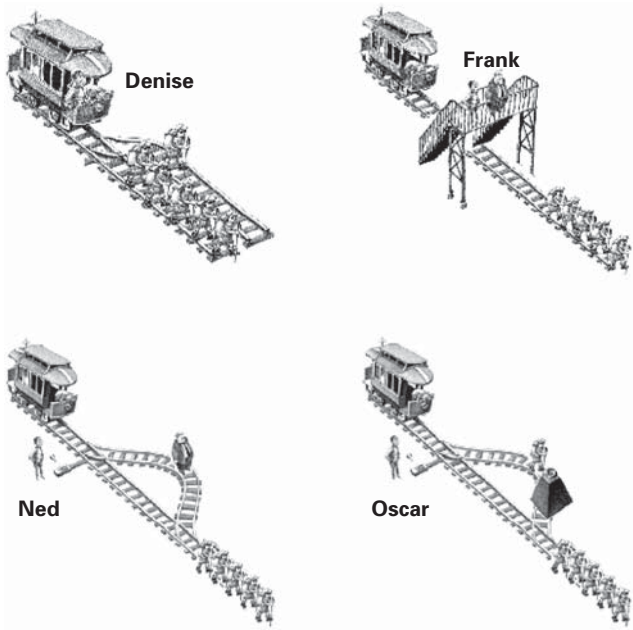

Figure 3.6
The core family of trolley dilemmas used in Internet studies of moral judgments and justifications.

can turn the trolley, killing the one; or she can refrain from turning the trolley, letting the five die.

Is it morally permissible for Denise to switch the trolley to the side track?

Scenario 2 Frank is on a footbridge over the trolley tracks. He knows trolleys and can see that the one approaching the bridge is out of control. On the track under the bridge there are five people; the banks are so steep that they will not be able to get off the track in time. Frank knows that the only way to stop an out-of-control trolley is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a large man wearing a backpack, also watching the trolley from the footbridge. Frank can shove the man with the backpack onto the track in the path of the trolley, killing him; or he can refrain from doing this, letting the five die.

Is it morally permissible for Frank to shove the man?

Scenario 3 Ned is taking his daily walks near the trolley tracks when he notices that the trolley that is approaching is out of control. Ned sees what has happened: the driver of the trolley saw five men walking across the tracks and slammed on the brakes, but the brakes failed and they will not be able to get off the tracks in time. Fortunately, Ned is standing next to a switch, which he can throw, that will tem-

porarily turn the trolley onto a side track. There is a heavy object on the side track. If the trolley hits the object, the object will slow the trolley down, thereby giving the men time to escape. Unfortunately, the heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the trolley from killing the men, but killing the man. Or he can refrain from doing this, letting the five die.

Is it morally permissible for Ned to throw the switch?

Scenario 4 Oscar is taking his daily walk near the trolley tracks when he notices that the trolley that is approaching is out of control. Oscar sees what has happened: the driver of the trolley saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The trolley is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Oscar is standing next to a switch, which he can throw, that will temporarily turn the trolley onto a side track. There is a heavy object on the side track. If the trolley hits the object, the object will slow the trolley down, thereby giving the men time to escape. Unfortunately, there is a man standing on the side track in front of the heavy object, with his back turned. Oscar can throw the switch, preventing the trolley from killing the men, but killing the man. Or he can refrain from doing this, letting the five die.

Is it morally permissible for Oscar to throw the switch?

As discussed in the philosophical literature, these cases generate different intuitions concerning permissibility. For example, most agree that Denise and Oscar are permissible, Frank is certainly not, and Ned is most likely not. What is problematic about this variation is that pure deontological rules such as "Killing is impermissible" or utilitarian considerations such as "Maximize the overall good" can't explain philosophical intuition. What might account for the differences between these cases? From 2003–2004—the first year of our project—over 30,000 subjects from 120 countries logged on to our Web site. For the family of four trolley dilemmas, our initial data set included some 5,000 subjects, most of whom were from English-speaking countries (Hauser, Cushman, Young, Jin, & Mikhail, 2006). Results showed that 89% of these subjects judged Denise's action as permissible, whereas only 11% of subjects judged Frank's action as permissible. This is a highly significant difference, and perhaps surprising given our relatively heterogeneous sample, which included young and old (13–70 years), male and female, religious and atheist/agnostic, as well as various degrees of education.

Given the size of the effect observed at the level of the whole subject population (Cohen's $d = 2.068$), we had statistical power of .95 to detect a difference between the permissibility judgments of the two samples at the .05 level given 12 subjects. We then proceeded to break down our sample

along several demographic dimensions. When the resultant groups contained more than 12 subjects, we tested for a difference in permissibility score between the two scenarios. This procedure asks: can we find any demographic subset for which the scenarios Frank and Denise do not produce contrasting judgments? For our data set, the answer was "no." Across the demographic subsets for which our pooled effect predicted a sufficiently large sample size, the effect was detected at $p < .05$ in every case but one: subjects who indicated Ireland as their national affiliation (see table 3.1). In the case of Ireland the effect was marginally significant at $p = .07$ with a sample size of 16 subjects. Given our findings on subjects' judgments, the principled reasoning view would predict that these would be accompanied by coherent and sufficient justifications. We asked subjects perceiving a difference between Frank and Denise to justify their responses. We classified justifications into three categories: (1) sufficient, (2) insufficient, and (3) discounted.

A sufficient justification was one that correctly identified any factual difference between the two scenarios and claimed the difference to be the basis of moral judgment. We adopted this extremely liberal criterion so as

Table 3.1
Demographic subsets revealing a difference for Frank vs. Denise

National Affiliation	**Religion**	**Education**
Australia	Buddhist	Elementary school
Brazil	Catholic	Middle school
Canada	Christian Orthodox	High school
Finland	Protestant	Some college
France	Jewish	BA
Germany	Muslim	Masters
India	Hindu	PhD
Ireland (p = .07)	None	**Ethnicity**
Israel	**Age**	American Indian
The Netherlands	10–19 yrs	Asian
New Zealand	20–29	Black non-Hispanic
Philippines	30–39	Hispanic
Singapore	40–49	White non-Hispanic
South Africa	50–59	**Gender**
Spain	60–69	Male
Sweden	70–79	Female
United States	80–89	
United Kingdom		

not to prejudge what, for any given individual, counts as a morally relevant distinction; in evaluating the merits of some justifications, we find it clear that some distinctions (e.g., the agent's gender) do not carry any explanatory weight. Typical justifications were as follows: (1) for Denise, the death of one person on the side track is not a necessary means to saving the five, while in Frank, the death of one person is a necessary means to saving the five; (2) in Denise, an existing threat (of the trolley) is redirected, while in Frank, a new threat (of being pushed off the bridge) is introduced; (3) in Denise, the action (flipping the switch) is impersonal, while in Frank, the action (pushing the man) is personal or emotionally salient.

An insufficient justification—category 2—was one that failed to identify a factual difference between the two scenarios. Insufficient justifications typically fell into one of three subcategories. First, subjects explicitly expressed an inability to account for their contrasting judgments by offering statements such as "I don't know how to explain it," "It just seemed reasonable," "It struck me that way," and "It was a gut feeling." Second, subjects explained that death or killing is "inevitable" in one case but not in the other without offering any further explanation of how they reasoned this to be the case. Third, subjects explained their judgment of one case using utilitarian reasoning (maximizing the greater good) and their judgment of the other using deontological reasoning (acts can be objectively identified as good or bad) without resolving their conflicting responses to the two cases. Subjects using utilitarian reasoning referred to numbers (e.g., save five vs. one or choose "the lesser of two evils"). Subjects using deontological reasoning referred to principles, or moral absolutes, such as (1) killing is wrong, (2) playing God, or deciding who lives and who dies, is wrong, and (3) the moral significance of not harming trumps the moral significance of providing aid.[7]

Discounted responses—category 3—were either blank or included added assumptions. Examples of assumptions included the following: (1) people walking along the tracks are reckless, while people working on the track are responsible, (2) a man's body cannot stop a trolley, (3) the five people will be able to hear the trolley approaching and escape in time, and (4) a third option for action such as self-sacrifice exists and should be considered.

When contrasting Denise and Frank, only 30% of subjects provided sufficient justifications. The sufficiency of subjects' justifications was not predicted by their age, gender, or religious background; however, subjects with a background in moral philosophy were more likely to provide sufficient justifications than those without.

In characterizing the possible differences between Denise and Frank, one could enumerate several possible factors including redirected versus introduced threat, a personal versus impersonal act, and harming one as a means versus a by-product. It is possible, therefore, that due to the variety of possible factors, subjects were confused by these contrasting cases, making it difficult to derive a coherent and principled justification. To address this possibility, we turn to scenarios 3 and 4—Ned and Oscar.

These cases emerged within the philosophical literature (Fischer & Ravizza, 1992; Kamm, 1998a; Mikhail, 2000) in order to reduce the number of relevant parameters or distinctions to potentially only one: means versus by-products. Ned is like Frank, in that a bystander has the option of using a person as the means to saving five. The person on the loop is a necessary means to saving the five since removing him from the loop leaves the bystander with no meaningful options: flipping the switch does not remove the threat to the five. The man on the loop is heavy enough to slow the trolley down before hitting the five. In Oscar, the man on the loop isn't heavy enough to slow the trolley, but the weight in front of him is. The weight, but not the man, is therefore a sufficient means to stopping the trolley. In both Ned and Oscar, the act—flipping a switch—is impersonal; consequently, on the view that Greene holds (Model 3), these should be perceived as the same. In both scenarios, the act results in redirecting threat. In both, the act results in killing one. In both, action is intended to bring about the greater good. But in Ned, the negative consequence—killing one—is the means to the positive—saving five—whereas in Oscar, the negative consequence is a by-product of a prior goal—to run the trolley into the weight so that it will slow down and stop before the five people up ahead.

Do subjects perceive these distinctions? In terms of judgments, 55% of subjects responded that it is permissible for Ned to flip the switch, whereas 72% responded that it is permissible for Oscar to flip the switch. This is a highly significant difference.

Paralleling our analysis of Frank and Denise, we calculated the necessary sample size to detect a difference between the cases assuming an effect size equal to the effect size of the total subject population (Cohen's $d = 0.3219$). Because of the substantially smaller effect size, a sample of 420 subjects was necessary to achieve statistical power of .95. Employing this stringent criterion, we were able to test a small range of demographic subsets for the predicted dissociation in judgments: males, females, subjects ages 30–39, 40–49, or 50–59, subjects who had completed college and subjects currently enrolled in college, Protestants and subjects indicating no religious

affiliation. For every one of these groups, the predicted dissociation in judgments was observed. In order to broaden the cross-cultural sample, we then tested additional demographic subsets for which we predicted statistical power of .8 to pick up a true effect. Again, every group showed the predicted dissociation in judgments. The additional groups were subjects ages 20–29 and 60–69, subjects who had completed high school but not enrolled in college, and Catholics.

Given that the Ned and Oscar cases greatly curtail the number of possible parametric differences, one might expect subjects to uncover the key difference and provide a sufficient justification. In parallel with Denise and Frank, only 13% of subjects provided a sufficient justification, using something like the means/by-product distinction as a core property.

Results from our family of trolley problems leave us with two conclusions: there is a small and inconsistent effect of cultural and experiential factors on people's moral judgments, and there is a dissociation between judgment and justification, suggesting that intuition as opposed to principled reasoning guides judgment. These results, though focused on a limited class of dilemmas, generate several interim conclusions and set up the next phase of research questions.

Consider first our four toy models concerning the causes of our moral judgments. If model 1—and its instantiation in the Kantian creature—provides a correct characterization, then we would have expected subjects to generate sufficient justifications for their judgments. Since they did not, there are at least two possible explanations. The first is that something about our task failed to elicit principled and sufficient explanations. Perhaps subjects didn't understand the task, didn't take it seriously, or felt rushed. We think these accounts are unlikely for several reasons. With few exceptions, our analyses revealed that subjects were serious about these problems, answering them as best as they could. It is also unlikely that subjects felt rushed given that they were replying on the Internet and were given as much time as they needed to answer. It is of course possible that if we had handed each subject a range of possible justifications that they would have arrived at the correct one. However, given their choice, we would not be able to distinguish between a principle that was truly responsible for their judgment as opposed to a post hoc rationalization. As Haidt has argued in the context of an emotionally mediated intuitive model, people often use a rational and reasoned approach as a way to justify an answer delivered intuitively. The second possibility, consistent with the Rawlsian creature, is that subjects decide what is permissible, obligatory, or forbidden based on unconscious and inaccessible principles. The reason

why we observed a dissociation between judgment and justification is that subjects lack access to the reasons—the principles that make up the universal moral grammar.

Our results, especially the fact that some subjects tended to see a difference between Ned and Oscar, also generates difficulties for both models 2 and 3. For subjects who see a difference between these cases, the difference is unlikely to be emotional, at least in the kind of straightforward way that Greene suggests in terms of his personal–impersonal distinction.[8] Both Ned and Oscar are faced with an action that is impersonal: flipping a switch. If Ned and Oscar act, they flip a switch, causing the trolley to switch tracks onto the loop, killing one person in each case but saving five. For Ned, the action of flipping a switch isn't bad. Flipping a switch so that the trolley can hit the man constitutes an action that can be more neutrally translated as "using a means to an end." If the heavy man had not been on the track, Ned would have no functionally meaningful options: flipping the switch, certainly an option in the strict sense, would serve no purpose as the trolley would loop around and hit the five people. In contrast, if the heavy man had not been on the looped track when Oscar confronted the dilemma, he could have still achieved his goal by flipping the switch and allowing the trolley to hit the heavy weight and then stop. The difference between Ned and Oscar thus boils down to a distinction between whether battery to one person was an intended means to saving five as opposed to a foreseen consequence. This distinction, often described as the "principle of double effect," highlights the centrality of looking at the causes and consequences of an action and how these components feed into our moral judgments.

The results discussed thus far lead, we think, to the intriguing possibility that *some* forms of moral judgment are universal and mediated by unconscious and inaccessible principles. They leave open many other questions that might never have been raised had it not been for an explicit formulation of the linguistic analogy, and a contrast between the four toy models and their psychological ingredients. For example, why are some moral judgments relatively immune to cross-cultural variation? Are certain principles and parameters universally expressed because they represent statistical regularities of the environment, social problems that have recurred over the millennia and thus been selected for due to their consistent and positive effects on survival and reproduction? Is something like the principle of double effect at the right level of psychological abstraction, or does the moral faculty operate over more abstract and currently unimaginable computations? Even though people may not be able to retrieve sufficient

justifications for some of their judgments, do these principles enter into future judgments once we become aware of them? Do results like these lead to any specific predictions with respect to the moral organ—the circuitry involved in computing whether an action is permissible, obligatory, or forbidden? In the next section, we describe a suite of ongoing research projects designed to begin answering these questions.

Universality, Dilemmas, and the Moral Organ

The Web-based studies we have conducted thus far are limited in a number of ways. Most importantly, they are restricted to people who not only have access to the Web and know how to use it but are also largely from English-speaking countries. Early Web-based studies were criticized for being uncontrolled and unreliable. These criticisms have been addressed in several ways. First, a number of experimental psychologists such as Baron and Banaji (Baron & Siepmann, 2000; Greenwald, Nosek, & Banaji, 2003; Kraut, Olson, Banaji, Bruckman, Cohen, & Cooper, 2004; Schmidt, 1997) have systematically contrasted data collected on the Web with data collected using more standard paper-and-pencil tests in a room with an experimenter. In every case, the pattern of results is identical. Similarly, our results on the Web are virtually identical to those that Mikhail and colleagues (1998) collected with the same dilemmas, but using paper-and-pencil questionnaires. Second, in looking over our data sets, we are rarely forced to throw out data from subjects who produce obviously faulty data, such as entering graduate degrees in the early teen years or linking nationality to the Antarctic. Third, for every test we administer on the Web, we include several control questions or dilemmas designed to test whether subjects understand the task and are taking it seriously.

In terms of cross-cultural diversity, we are currently stretching our reach in two different directions. First, we have already constructed translations of our Web site into Arabic, Indonesian, French, Portuguese, Chinese, Hebrew, and Spanish and have launched the Chinese and Spanish Web sites. Second, we have initiated a collaboration with several anthropologists, economists, and psychologists who are studying small-scale societies in different parts of the world. Under way is a study with Frank Marlowe designed to test whether the Hadza, a small and remote group of hunter-gatherers living in Tanzania, show similar patterns of responses as do our English-speaking, Internet-sophisticated, largely Westernized and industrialized subjects. This last project has forced us to extend the range of our dilemmas, especially since the Hadza, and most of the other small-scale

societies we hope to test, would be completely unfamiliar with trolleys. Instead of trolleys, therefore, we have mirrored the architecture of these problems but substituted herds of stampeding elephants as illustrated below (see figure 3.7). Like Denise, the man in the jeep has the option of watching the herd run over and kill five people or of driving toward the herd, turning them away from the five and around the grove where they will run over and kill one person. Similarly, in a case designed to mirror Frank, a person can throw a heavy person out of a tree to stop the herd and thereby save the five people up ahead. Marlowe's preliminary data suggest that the Hadza judge these cases as do Web-savvy Westerners and, also, fail to give sufficient justifications. Though preliminary, these results provide further support for the universality of some of our moral intuitions.

Changing the content of these dilemmas not only is relevant for testing small-scale societies that are unfamiliar with trolleys but also makes precisely the right move for extending the reach of our empirical tests. In

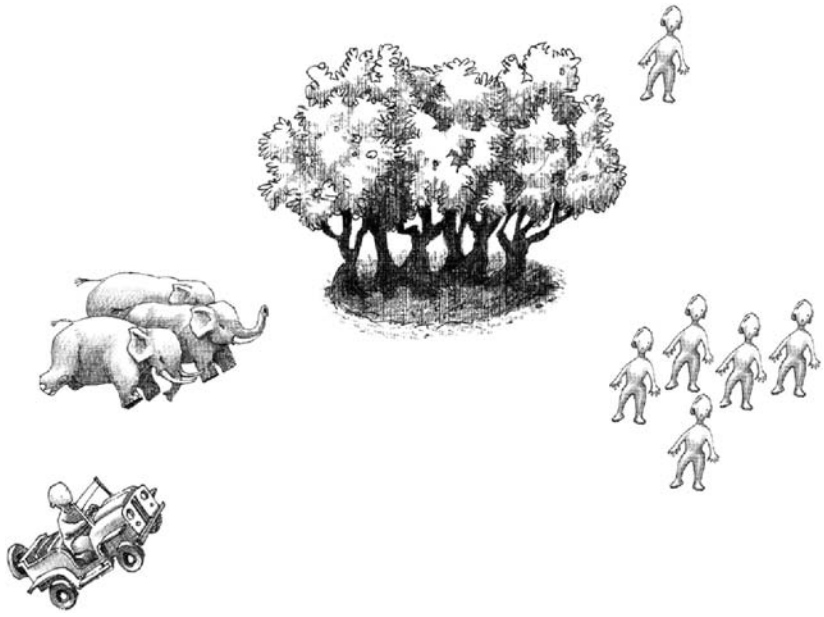

Figure 3.7
A content manipulation of the familiar bystander trolley problem, designed for field testing among hunter-gatherer populations. Here, a man in a jeep has an opportunity to drive toward the herd of stampeding elephants, causing them to move around the grove, saving the five but killing the one person.

particular, we have now constructed several hundred dilemmas, each carefully articulated in terms of the text, while systematically manipulating the content of the dilemma, the nature of the action, the consequences of action as opposed to inaction, the degree to which the consequences are a direct or indirect result of the action, and so forth. More specifically, we have mined the rich philosophical literature on moral dilemmas, including cases of harm, rescue, and distribution of limited resources, to derive a series of relevant parameters and potential principles for building a library of dilemmas that can be presented to subjects on the Web, in the field, and in hospital settings with patient populations.

The strongest opposition to the strict Kantian creature has been the Humean creature. And yet, as we have tried to argue throughout, it is not at all clear how our emotions play a role. As suggested in the first section, *that* emotions play a role is undebatable. To more precisely identify where, when, and how emotions play a role in our moral judgments, we have initiated a suite of collaborative projects with cognitive neuroscientists using patient populations with selective brain damage, functional neuroimaging, and transcranial magnetic stimulation. Here, we give only a brief sketch of some preliminary results and their potential significance for fleshing out the details of our moral psychology.

Over the past 15 or more years, Antonio Damasio (1994, 2000) has amassed an impressive body of data on the neurobiology of emotion and how it bears on our decision making. Some of the most intriguing results come from his studies of patients with damage to the orbitofrontal and ventromedial prefrontal cortex. Based on a wide variety of tests, it appears that these patients often make inappropriate decisions *because* of insufficient input from the emotional circuitry of the brain. This also leads to what appear to be inappropriate moral decisions. On the face of it, this might be taken as evidence for the Humean creature. In the absence of emotional input, moral judgments are often at odds with what nonpatients say. However, because there have been insufficient, in-depth tests of their moral psychology, it is not clear how extensive the deficit is, nor whether it is due to performance or competence. Given the lack of clarity, we teamed up with Damasio, Ralph Adolphs, Daniel Tranel, and Michael Koenigs (2007) and began testing these patients[9] on a large battery of moral dilemmas, including the original family of trolley problems, several additional permutations, and many other dilemmas aimed at different aspects of our moral psychology. For several dilemmas, these patients showed completely normal pattern of judgments. This shows that emotions are not necessary for a variety of moral situations. However, in cases where a

highly aversive action is in conflict with the generation of a significant utilitarian outcome, and the action involves personal contact with another, these patients deviate significantly from normals, favoring the utilitarian outcome. That is, in this selective set of moral problems, emotions appear causally necessary. When the circuitry subserving social emotions is damaged, a hyper-utilitarian emerges.

These results only skim the surface of possibilities and only present a rough picture of the different computations involved in both recognizing a moral dilemma and arriving at a judgment. Crucially, by laying out the possible theoretical issues in the form of our four toy models, and by taking advantage of empirical developments in cognitive neuroscience, we will soon be in a exquisite position to describe the nature of our moral judgments, how they are represented, and how they break down due to acquired or inherited deficits.

Sweet Justice! Rawls and Twenty-first-century Cognitive Science

In 1998, Rawls wrote *Justice as Fairness*, one of his last books. In some sense, it represents the finale to his work in political philosophy, providing the interested reader with an update on his thinking since 1971 when he published *A Theory of Justice*. For the observant reader, there is something missing in this final installment: the linguistic analogy has been completely purged! This is odd on at least two counts. First, linguistics as a discipline was stronger than it had ever been, and certainly in a far more mature state than it was in the 1970s. Not only had there been considerable theoretical developments but work in linguistics proper had joined forces with other neighboring disciplines to provide beautiful descriptions of the neural architecture and its breakdown, the patterns of development, the specificity of the machinery, and the historical and evolutionary patterns of change. Building the analogy would have been, if anything, easier in 1998 than it was at the time Rawls first began writing about language and morality; fortunately, other philosophers including Gert, Dwyer, and Mikhail have picked up where Rawls left off. Second, our understanding of cognitive processes more generally, and moral psychology more specifically, had grown considerably since Piaget and Kohlberg's writings between 1960 and 1980. In particular, many of the issues that Rawls was most deeply interested in concerning principles of justice qua fairness were being explored by political scientists and economists, in both developed and developing countries—an empirical march that continues today (Camerer, 2003; Frohlich & Oppenheimer, 1993; Henrich, Boyd, Bowles,

Camerer, Fehr, & Gintis, 2004). It is in part because of these developments that the time is ripe to bring them back and flesh out their empirical implications.

As stated earlier, there is a strong and weak version of the linguistic analogy. On the strong version, language and morality work in much the same way: dedicated and encapsulated machinery, innate principles that guide acquisition, distinctions between competence and performance, inaccessible and unconscious operative principles, selective breakdown due to damage to particular areas of the brain, and constraints on the evolvable and learnable languages and moralities.[10] On the weak version, the linguistic analogy is merely a heuristic for posing the right sorts of questions about the nature of our moral competence. On this version, it matters little whether morality works like language. What matters is that we ask about the principles that guide mature competence, work out how such knowledge is acquired, understand whether and how competence interacts with both mind internal and external factors to create variation in performance, and assess how such knowledge evolved and whether it has been specially designed for the moral sphere. These are large and important questions, and, to date, we have few answers for them.[11]

Providing answers will not be trivial, and for those interested in moral knowledge and the linguistic analogy in particular, one must recognize that the state of play is far worse than it was when Chomsky and other generative grammarians began writing about language in the 1950s. In particular, whereas linguists have been cataloguing the details of the world's languages, dissecting patterns of word order, agreement, and so on, we have nothing comparable in the moral domain. In the absence of a rich description of adult moral competence, we can't even begin to work out the complexity of the computations underlying our capacity to create and comprehend a limitless variety of morally meaningful actions and events. And without this level of descriptive adequacy, we can't move on to questions of explanatory adequacy, focused in particular on questions of the initial state of competence, interfaces with other mind internal and external factors, and issues of evolutionary uniqueness. On a positive note, however, by raising such questions and showing why they matter, we gain considerable traction on the kinds of data sets that we will need to collect. It is this traction that we find particularly exciting and encouraging in terms of working out the signature of our moral faculty.

Let us end on a note concerning descriptive as opposed to prescriptive ethics. Rawls's linguistic analogy is clearly targeted at the descriptive level, even though many of his critics considered him to be saying more

(Mikhail, 2000, in press). Showing how the descriptive level connects to the prescriptive is a well-worn and challenging path. Our own sense, simple as it may be, is that by understanding the descriptive level we will be in a stronger position to work out the prescriptive details. This is no more (or less) profound than saying that an understanding of human nature, how it evolved, and how it has changed over recent times provides a foundation for understanding our strengths and weaknesses and the kinds of prescriptive policies that may or may not rub up against our innate biases. As an illustration, consider the case of euthanasia and the distinction made by the American Medical Association (AMA) between mercy killing and removing life support. This example, well-known to moral philosophers (Kagan, 1988; Rachels, 1975), is precisely the kind of case that motivated the development of the trolley problems. It is an example that plays directly into the action versus inaction bias (Baron, 1998). The AMA blocks a doctor's ability to deliver an overdose to a patient with a terminal and insufferable illness but allows the doctor to remove life support (including the withdrawal of food and fluids), allowing the patient to die. The AMA allows passive euthanasia but blocks active euthanasia. Although this policy feeds into an inherent bias that we appear to have evolved in which actions are perceived as more harmful than inactions, even when they lead to the same consequences, it is clear that many in the medical community find the distinction meaningless. The intuition that the distinction is meaningless appears even stronger in a different context: James Rachels's example of a greedy uncle who intends to end his nephew's life in order to inherit the family's money, and in one case drowns him in the bathtub and in another lets him drown. His intent is the same in both cases, and the consequences are the same as well. Intuitively, we don't want to let the uncle off in the second case, but convict him of a crime in the first. And the intuition seems to be the same among medical practitioners. Indications that this is the case come from several lines of evidence, including the relatively high rate of unreported (and illegal!) mercy killings going on every day in hospitals in the United States, the fact that many patients diagnosed with some terminal illness often "die" within 24 hours of the diagnosis, and the fact that some countries, such as The Netherlands and Belgium, have abandoned the distinction between active and passive euthanasia altogether. All in all, intuition among medical practitioners appears to go against medical policy.

The fact that intuition rides against policy doesn't mean, in general, that we should allow intuition to have its way in all cases. As Jonathan Baron and others have pointed out, intuition often flies in the face of what ulti-

mately and rationally works out to be the better policy from the standpoint of human welfare. However, ignoring intuition altogether, and going for rational deliberate reasoning instead, is also a mistake. Providing a deeper understanding of the nature of our intuitive judgments, including the principles that underlie them, how they evolved, how they develop, and the extent to which they are immune to reason and unchangeable, will only serve to enhance our prescriptive policies.

The issue of immunity or penetrability of our intuitive system brings us back to Rawls, and perhaps the most significant difference between language and morality. Looking at the current landscape of research in linguistics makes it clear that the principles underlying adult competence are phenomenally complex, abstract, and inaccessible to conscious awareness. The fact that those studying these principles understand them and have access to them doesn't have any significant impact on their performance, or what they use such principles for in their day-to-day life, from writing and reading to giving lectures and schmoozing at a café or pub. On the other hand, our strong hunch is that once we begin to uncover some of the principles underlying our moral judgments, they most certainly will impact our behavior. Although the principle of double effect may not be at the right level of abstraction, it is the kind of principle that, once we are aware of it, may indeed change how we behave or how we perceive and judge the behavior of others. In this sense, our moral faculty may lack the kind of encapsulation that is a signature feature of the language faculty. This wouldn't diminish the usefulness of Rawls's linguistic analogy. Rather, it would reveal important differences between these domains of knowledge and serve to fuel additional research into the nature of the underlying mechanisms, especially the relationship between competence and performance, operative and expressed principles, and so on. In either case, it would entail a gift to Rawls's deep insight about the nature of our moral psychology, an instance of sweet justice.

Notes

We would like to extend our deepest thanks to John Mikhail for helping us to clarify many of the links between language and morality, and rebuilding Rawls's analogy. Also, thanks to Walter Sinnott-Armstrong for organizing a terrific conference among philosophers, biologists, and psychologists, and for giving us extensive comments on this paper; thanks too to three undergraduates in his class for helping us clarify the issues for a more general readership. Hopefully, none of our commentators will be too disappointed. Mark Hauser was funded by a Guggenheim award during the writing of this chapter.

1. The description of judgments is not meant to exclude others including what is virtuous, ideal, and indecent. Throughout, however, we refer to judgments of permissible, obligatory, and forbidden actions, largely because these are the ones that we have focused on empirically. However, the Rawlsian theory that we favor will ultimately have to encompass judgments of actions that are morally right or wrong, good or bad, and above and beyond the call of duty. In parallel, most of the examples we will target concern harming. However, if the theory is to have sufficiently broad appeal, it will have to encompass harmless acts that are treated as moral infractions. For example, many of the dilemmas that we are currently exploring concern cases of rescue and resource contributions to those in need, as well as actions that are treated as morally impermissible because they are disgusting. It is too early to say whether the Rawlsian view we favor can do the work necessary to account for these other cases, but our hunch is that it will.

2. Rawls' views on the linguistic analogy are presented in section 9 of *A Theory of Justice*, but the precursor to this discussion originates in his thesis and the several papers that followed. For example, in his thesis he states, "The meaning of explication may be stated another way: ordinarily the use of elaborate concepts is intuitive and spontaneous, and therefore like 'cause,' 'event,' 'good,' are applied intuitively or by habit, and not by consciously applied rules. . . . Sometimes, instead of using the term 'explication' one can use the phrase 'rational reconstruction' and one can say that a concept is rationally reconstructed whenever the correct rules are stated which enable one to understand and explain all the actual occasions of its use" (pp. 72–73). Further on, he states that moral principles are "analogous to functions. Functions, as rules applied to a number, yield another number. The principles, when applied to a situation yield a moral rule. The rules of common sense morality are examples of such secondary moral rules" (p. 107). See Mikhail (2000) for a more comprehensive discussion of Rawls's linguistic analogy, together with several important extensions.

3. Our characterization of the Kantian creature is completely at odds with Greene's characterization. For Greene, whose ideas are generally encapsulated by Model 3, Kant is aligned with deontological views and these are seen as emotional. Although we think this is at odds with Kant, and others who have further articulated and studied his ideas, we note here our conflict with Greene's views.

4. Throughout the rest of this paper, when we use the terms "right," "wrong," "permissible," and so forth, we are using these as shorthand for "morally right," "morally wrong," morally "permissible," and so forth.

5. We should note that, as it is for Kamm, this is a methodological move. The moral faculty presumably handles real-world cases in the same way; the problem is that it may be more difficult to separate out competence–performance issues when it comes to real-world problems where people have already decided.

6. There are many permutations of these trolley problems, and in our research we have played around with framing effects (e.g., using "saving" as opposed to "killing"), the location of a bystander (e.g., Denise is on the trolley as opposed to on the side, next to the switch), and so on; in general, these seem to have small effects on overall judgments as long as the wording is held constant across a set of different dilemmas (e.g., if a permissibility question is framed with "saving," then all contrasting dilemmas use "saving" as well).

7. Our analyses of justifications are only at the crudest stage and may blur distinctions that certain subjects hold but do not make explicit. For example, subjects who justify their answers by saying that killing is wrong may have a more nuanced view concerning cause and effect, seeing Denise as carrying out an act that doesn't kill someone, whereas Frank's act clearly does. At present, we take the methodologically simpler view, using what people said as opposed to probing further on the particular meanings they assigned to different pieces of the justification.

8. It is possible that a different take on emotional processing could be used to account for the difference between Ned and Oscar; for example, as Sinnott-Armstrong suggested to us, a difference between imagining the victim jumping off the track in Ned frustrates our attempt to stop the trolley, which may be negatively coded, whereas the same event in Oscar would make things easier, and may be positively coded.

9. At present, we have tested six patients with frontal damage. The extent and location of damage is quite similar across patients.

10. Though not addressed explicitly in this paper, it is important to distinguish—as Chomsky has—between the internal computations underlying language and morality [I-language and I-morality] and the external representations of these computations in the form of specific E-languages (Korean, English, French) and E-moralities (permissible infanticide, polygyny).

11. Since the writing of this chapter in 2005, most of the references to our own work have changed from in preparation to in print, and dozens of other papers by our colleagues have emerged, transforming this rich landscape.

3.1 Reviving Rawls's Linguistic Analogy Inside and Out

Ron Mallon

Marc Hauser, Liane Young, and Fiery Cushman's paper is an excellent contribution to a now resurgent attempt (Dwyer, 1999; Harman, 1999; Mikhail, 2000) to explore and understand moral psychology by way of an analogy with Noam Chomsky's pathbreaking work in linguistics, famously suggested by John Rawls (1971). And anyone who reads their paper ought to be convinced that research into our innate moral endowment is a plausible and worthwhile research program. I thus begin by agreeing that even if the linguistic analogy turns out to be weak, it can do titanic work in serving "as an important guide to empirical research, opening doors to theoretically distinctive questions that, to date, have few answers" (p. ••). Granting the importance of the empirical investigation of moral judgment generally, and of research designed to probe the linguistic analogy specifically, I will nonetheless argue that there is simply no evidence that there is a specialized moral faculty, no evidence that the stronger version of the linguistic analogy is correct (p. ••).

What Is the Moral Faculty?

On the strong version of the linguistic analogy, Hauser et al. suggest that the moral faculty may be

1. A specialized system.
2. Innate.
3. Universal (i.e., species-typical).
4. Upstream of moral judgment (weak processing view).
5. Causally responsible for moral judgment, independent of emotion and reasoning (strong processing view).[1]

I am inclined to agree with them that there must be action appraisals that are upstream of moral judgment and that the capacity for such appraisals

may well be substantially innate. However, I am skeptical that there is any evidence that such appraisals involve a specialized moral faculty (1).

What do Hauser et al. mean by suggesting that there is a specific moral faculty? While they do not say a lot about what would make the moral faculty specialized, what I think they have in mind here is that there is a distinct mental subsystem that

- Properly functions in the domain of morality.
- Is functionally (computationally) discrete in that it makes use of only limited sorts of information (i.e., it exhibits information encapsulation) and in that its operational principles and processes are opaque to conscious reasoning.
- Is physiologically discrete—it is some sort of "organ" with a particular brain location.

These commitments lead them to talk of "specialized" moral systems (p. ••), "dedicated and encapsulated machinery" that exhibits "selective breakdown due to damage to particular areas of the brain" (p. ••). Before going on to argue against such a faculty, let's pause to ask what exactly they take the moral faculty to do.

Hauser et al. suggest the moral faculty is an "appraisal system" (p. ••) for "action analysis" (p. ••). Somewhat puzzlingly, they then go on to explain that actions can be combined to create events (p. •• ff). This is puzzling because the appraisal system is not the system that produces actions, but the one that appraises them. The idea that I think they have in mind is that actions have a complex structure that is isomorphic to the complex computational description of the action that we assign to the action prior to moral judgment. If I understand them correctly, then, there are at least two functions involved in what they call "action analysis." The first is the (perhaps automatic and unconscious) assignment of a description to an action, perhaps one that "computes the cause and consequences" of the action (p. ••). The second is the application of moral principles to such a description to result (directly or indirectly) in moral judgment.[2]

It seems to me that it must be the second of these functions that Hauser et al. want to identify with the moral faculty, for, at first look anyway, the assignment of a description to an action does not look to be specific to the moral domain nor to be informationally encapsulated. It seems we can, for example, assign action descriptions to actions toward which we have no moral reaction. Nor does it seem that moral principles need be involved in such an assignment. Moreover, it seems that such action descriptions

are assigned using information from a wide variety of sources (e.g., general knowledge, theory of mind, etc.). This is not a problem for Hauser et al., because the second function, the application of principles or rules to actions resulting in moral judgment, may be functionally discrete. Thus, if there is a specialized moral faculty, it is the computational mechanism that takes action descriptions as inputs, applies moral principles and parameters, and gives moral judgments (or, on the weak processing view, precursors to moral judgment like moral reasoning or emotion) as outputs.

Three Projects in Understanding Moral Judgment

Let's distinguish three different projects in the area of moral judgment: the first one prescriptive, the other two descriptive. First, many people in general, and moral philosophers in particular, are typically interested in what the correct moral assessment of a particular sort of person or act is. The correct *prescriptive* account of moral judgment (if there is one) could, in principle, allow us to understand for any object (e.g., a person or an action) what the appropriate moral evaluation of that object is. It would perhaps tell us that murder is wrong or keeping one's promises is right, but knowing such a theory would also enable us to know the right answer (if there is one) in hard cases like the moral dilemmas many philosophers focus on.

Of course, moral psychologists like Hauser at al. are not primarily interested in the prescriptive project. Rather, they are interesting in a descriptive project of accurately characterizing the capacities that give rise to moral judgment. However, as Shaun Nichols (2005) has recently pointed out (in a similar context), this descriptive project admits of *external* and *internal* readings similar to those that arose in discussions of linguistics. On the *external* view of the linguistic project, a primary aim of the linguist is to produce a descriptively adequate grammar that predicts linguistic intuitions of speakers and is consistent with the developmental and cross-cultural data. Choice of such a grammar might be further constrained by other theoretical considerations such as simplicity, but crucially, such an adequate grammar could well have principles quite at odds with anything that is subserved by a specific mechanism or actually represented in language users. On the externalist view, a gap between the principles our theory invokes and the psychological mechanisms that subserve the processes our theory describes is perfectly okay, for the external project is psychologically modest (e.g., Stich, 1972). In contrast, the *internal* reading holds that the project of linguistics is to describe the psychological

mechanisms (perhaps including the principles and parameters) that actually give rise to—in virtue, perhaps, of their being mentally represented—judgments of grammaticality in a mature, competent native speaker. The internal project is thus psychologically ambitious: it aims, inter alia, to provide a description of the computational mechanisms that instantiate the adult native speaker's competence with language (e.g., Fodor, 1981).

The same distinction can be applied to the project of understanding adult moral capacity. Here, the external project would be to characterize a descriptively adequate set of moral principles that capture our moral judgments, including those regarding trolley cases. Such a project need not be committed to moral principles that can be explicitly articulated but rather can include whatever principles seem to capture and order the relevant intuitive judgments. In contrast, the more psychologically ambitious internal project aims to characterize those computational mechanisms that actually give rise to our moral judgments.

Rereading Rawls, Outside In

Hauser et al. are pursuing the *internal* project of characterizing the underlying mechanisms that explain moral judgment, and they postulate a moral faculty as part of this project. A critic can thus allow them that there are (externally adequate) moral rules that are innate, are universal (i.e., species typical), and figure in the production of moral judgment while nonetheless denying that there is a functionally discrete faculty that computes from action descriptions to moral judgments.[3]

Before I go on to make that argument, however, it is worth noting that Rawls himself seems most plausibly read as interested in the external project. When Rawls (1971) writes, "A correct account of moral capacities will certainly involve principles and theoretical constructions which go beyond the norms and standards cited in every day life" (p. 47), he is not merely indicating that our moral capacities may involve moral principles that go beyond those we can express, as Hauser et al. suggest. Rather, Rawls seems to be noting that the correct description of our moral capacity may outrun what can plausibly be literally attributed to the individual's psychological endowments. Rawls's aim in the passage Hauser et al. cite is, in part, to defend the relevance of his device of the "original position"—a theoretical construction that we ought not regard him as holding to be an actual component (conscious or unconscious) of our processing of moral judgments. While Hauser et al.'s passage from Rawls ends with, "A correct account of moral capacities will certainly involve principles and theoreti-

cal constructions which go beyond the norms and standards cited in everyday life" (p. ••), the original sentence and paragraph continue as follows:

> it may eventually require fairly sophisticated mathematics as well. This is to be expected, since on the contract view the theory of justice is part of the theory of rational choice. Thus the idea of the original position and of an agreement on principles there does not seem too complicated or unnecessary. Indeed, these notions are rather simple and can serve only as a beginning. (Rawls, 1971, p. 47)

The full passage offers just an inkling of how much theoretical apparatus Rawls thinks may appropriately be invoked in the course of characterizing our moral capacity, apparatus that seems unconstrained by the psychological facts of processing.

It would, of course, be a mistake to assume that Hauser et al.'s use of the linguistic analogy stands or falls with successful Rawls exegesis. However, once we distinguish the external and internal projects, it does raise questions about how the two projects, if both carried out, might relate to one another. In particular, it might well be that an external theory could be developed, setting out principles governing moral judgment within a moral culture and parameters that vary among moral cultures, but that the principles of such a theory are not smoothly reducible to specific principles or a specific faculty operative in psychological processing.

Prying Apart the External and the Internal Projects

It is something of a truism in cognitive science that functional identified domains like "moral judgment" may be numerously instantiated computationally, so that there is no reason to infer from the seeming coherence of the folk category "moral judgment" that the psychological mechanisms producing such judgments will themselves cohere. Here, I will develop that idea.

Multiple Realizability

A central organizing doctrine of much cognitive science is that cognitive behavioral phenomena can be described at multiple levels (see, e.g., Marr, 1982; Newell, 1982; Pylyshyn, 1984). A typical division of levels of description might involve three levels:

1. A descriptive level: Describes the function performed by the target mechanism.
2. A computational level: Describes the algorithm actually used to compute the function described in (1).

3. Implementation level: Describes the physical materials that implement the computation described in (2).

Within this tradition, one can think of the external descriptive project as offering a high-level description of principles for judgment that adequately characterize the functional domain (1), while the internal project (2) attempts to specify the computational mechanisms that implement or realize those higher level principles.[4]

Most simply, we can imagine a correct and complete description of our moral capacity (Level 1) invokes simple moral principles that are literally represented in the brain and used in computations to generate moral judgment (Level 2), and this is all carried out in a particular way by a particular, functionally distinct brain region (Level 3). However, an equally familiar point from these discussions in cognitive science is that properties functionally specified at a higher level of description may be realized by a variety of different lower level mechanisms. Thus the mere fact that we can describe specific jobs for a moral faculty (e.g., action appraisal) ought to give us no confidence at all that there really is some specialized computational faculty (Level 2) or brain region (Level 3) that realizes such a function. Rather, it might well be that multiple or diffuse internal mechanisms operate in such a way that we can accurately describe them (at Level 1) as performing (or computing) the function. The mere possibility of such multiple realizability ought to undermine any easy faith that a principle or set of principles operating in an adequate description of our moral capacities will find smooth reduction to particular psychological mechanisms. Given an adequate description of our moral capacity, there are just too many underlying computational architectures that could play such a role.

Nichols on Double Effect

Nichols (2005) has recently made just this point in just this context, so let me rehearse his idea, and then discuss its implications for Hauser's account. Nichols's discussion begins with Gilbert Harman's (1999) suggestion that the doctrine of double effect (DDE) might be "built into" people, forming part of our "universal moral grammar" (p. 114). Nichols goes on to point out that even if the DDE "is externally adequate to a core set of Trolley intuitions, we still need to determine the best internal account" and "it is by no means clear that the appeal to an innate DDE principle is the best explanation" (p. 361). Nichols points out that the DDE includes multiple criteria for assessing the permissibility of an action, for example, it includes both of the following:

1. The requirement that the good outcome of the action be greater than the bad outcome.
2. The requirement that the bad outcome not be intended.

Given such a complex principle, Nichols goes on to sketch how it might be that different faculties may underlie distinct criteria, suggesting that a system for utilitarian calculations may underlie (1) while a distinct system (what he calls "deontological system") may underlie (2).

Now Hauser et al. do not have much faith that the DDE is a principle of universal moral grammar, perhaps because it is not complex and abstract enough (p. ••). But Nichols's point here is entirely generalizable: the mere fact that we can describe principles that seem to capture intuitions about a set of moral cases gives us exactly no reason at all to think that those principles are themselves implemented directly in a computationally discrete way or by a computationally discrete faculty.

A Yawning Gap: How External and Internal Projects May Have Divergent Aims

Hauser and his colleagues invoke the venerable "performance–competence" distinction, but now that we have distinguished internal and external linguistic projects, we can draw this distinction for either project. A natural reading of this distinction on an internal approach is to say that in seeking to characterize a competency, we aim to literally specify the distinct organizations of the various computational mechanisms that constitute a mind. The distinction between competence and performance is just a way of indicating that while our behavioral evidence typically results from a combination of factors, we are trying to draw inferences about *computational competence*—about the computational structure of a particular module or mechanism.

However, even an external approach to morality must make use of such a distinction, for here too, the theorists will be faced with distinguishing data that genuinely reveal moral considerations from those that do not. Rawls (1971), for example, privileges "considered judgments" as

> those judgments in which our moral capacities are most likely to be displayed without distortion. . . . [Judgments] given when we are upset or frightened, or when we stand to gain one way or the other can be left aside. All these judgments are likely to be erroneous or be influenced by an excessive attention to our own interests. . . . relevant judgments are those given under conditions favorable for deliberation and judgment in general. (pp. 47–48)

In thinking about the principles underlying our moral capacity, Rawls seeks something we can call "domain competence," and he would have us put our "considered judgment" at the core of our enterprise. We can leave it an open question as to whether he is right, and also to what extent the kind of data Rawls considers relevant is like the data that Hauser and his colleagues rely upon. Instead, we simply note that a project focused on domain competence might hold very different judgments to be relevant than one focused on computational competence.

To see this, recall that Hauser and his colleagues emphasize the importance of evolution in thinking about the structure of our moral faculty (p. ••), but they curiously assign evolution little role in determining our computational competence. For example, Petrinovich, O'Neill, and Jorgensen (1993) report finding that subjects prefer the lives of relatives and friends over strangers in standard trolley scenarios, a finding they take to support sociobiologists' and evolutionary psychologists' suggestions that humans are designed, in part, to be concerned with their own inclusive fitness. Hauser et al. indicate that in contrast with such research that focuses on questions of "evolutionary significance," their research will probe "the computational operations that drive our judgments" (p. ••). However, this begs a crucial question, namely, whether the computational process driving our typical moral judgments are themselves biased by evolution in ways that are at odds with domain competence. Suppose that the data Petrivonich et al. report are correct, and moreover, suppose that much of our moral judgment is underwritten by an evolutionarily designed mechanism M that computes using the following internalized principle:

(K) The wrongness of a death is inversely proportional to the subject's relatedness to me.

The question is, would such a principle be part of the moral faculty Hauser et al. posit? On an external investigation into our domain competence, the answer might well be "no," for the biasing of judgments toward relatives might be thought of as a distortion of, rather than a part of, moral judgment. This seems to be Rawls's view. In contrast, the internal, psychologically ambitious project ought to want to understand mechanism M however it works, and whether or not we would want to say that its computation is part of morality.[5] Notice that from the internal point of view, to consider the computational function of M in cases (e.g., trolley cases without relatives) where it does not employ K is precisely to fail to characterize its computational competence.

This all goes to show simply that the external conception of moral competence and the internal conception can, and likely do, diverge. For

example, if the evolutionary hypothesis we have been considering is true, judgments about relatives might be irrelevant to domain competence but central to computational competence. Insofar as our best account of our moral domain competence, and our best account of the computational competences of our various cognitive mechanisms fail to neatly align, to that extent it will be wrong to say a specialized faculty underlies our moral domain competence.

Looking to the Data

Of course, all these arguments about the way things might go will be worth nothing if the experimental data support the strong linguistic analogy. And here they look to be in a very strong position, for they do have a very impressive research program gathering data on moral dilemmas within and across cultures. They consider four kinds of data that I will review here: data regarding selective deficits, data regarding judgments about moral dilemmas, cross-cultural data on moral dilemmas, and data regarding justifications for those moral judgments. None of these, I argue, provide evidence for a specialized moral faculty.

With regard to the data on selective deficits that they mention in passing, such deficits would, if borne out, support the strong linguistic analogy, for they would show that whatever underlies our capacity for moral appraisal has at least some necessary components that are physically localized in the brain. Here I will only say (in agreement with Jesse Prinz's commentary on Hauser et al. in this volume), that there is, to my knowledge, simply no evidence at this time for selective deficits of a faculty that takes action descriptions as inputs and gives moral judgments (or their precursors) as outputs.

With regard to evidence of converging judgments on moral dilemmas, both within and across cultures, we should note that this sort of evidence is simply the wrong kind of evidence to bear on the question of whether there is a specialized moral faculty or whether the capacities to make these judgments are distributed throughout multiple different psychological mechanisms. These data *do* bear on the claim that *whatever mental faculties* underlie these judgments are innate and universal (i.e., species typical), but they do not give any evidence at all that there is one mental faculty rather than several, hundreds, or thousands. This is worth emphasizing: on an external approach to the data, one can describe shared judgments about moral features as revealing underlying shared principles and differences as resulting from diverse parameters, as a means of organizing the

data. But there's no reason at all to think such an organization reveals internal computational or physiological "joints" of the mind.

Finally, the data on justifications look to be either silent on whether, or undermine the case that, a single moral faculty is involved. The data on justifications are silent on whether there is a single moral faculty if one takes the justifications to be wholly unconnected with individual reasons for judgments. On this view, justifications are just post hoc rationalizations of one's prior judgments (Haidt, 2001). But if the justifications are wholly unconnected with the processing mechanisms, the content of the justifications provides no evidence for the features salient in the processing of the moral judgment. On the other hand, suppose that the justifications are based in part on introspective access to the reasons for actions. Then the fact that subjects who provided insufficient justifications sometimes appealed to diverse and unreconciled factors (Hauser et al., p. ••) seems to cohere precisely with Nichols's suggestion that there may be diverse and competing mechanisms in play in producing judgments about trolley cases.

In short, there is simply no evidence that supports positing a specialized moral faculty, and there is some that suggests just the opposite: our capacities in the moral domain result from the complex interactions of a variety of mental mechanisms not specific to the moral domain.

Conclusion

I have argued that there is no evidence to support the idea that there is a specialized moral faculty. In closing, I simply note how much this grants the linguistic analogy: implementing mechanisms for our moral capacities might well be innate, and they might even realize universal moral principles, modified by certain parameters, if this is understood as part of an external project. And yet, if the mechanisms themselves are not computationally discrete, if their computational competencies do not smoothly underlie domain competence in moral judgment, then the central idea of the strong linguistic analogy—the idea that there is a unified moral faculty—will simply be wrong. In its place might well be a messy mishmash of mental mechanisms that are not computationally of a piece.[6]

Notes

1. Hauser et al. distinguish between weak and strong processing roles for the moral faculty (pp. ••, ••) and also weak and strong versions of the linguistic analogy

(pp. ••, ••). My argument is against the strong version of the linguistic analogy, and against both processing views insofar as they endorse (1).

2. The computational economy they envisage (p. ••) arises because the principles employed by the moral faculty they envisage encompass an indefinitely large range of action descriptions.

3. Some might find it confusing to think that a moral rule could be innate but not "internal" in the sense described here. However, innateness typically involves a commitment to robust development across a wide variety of circumstances (Stich, 1975; Ariew, 1996; Sober, 1998) along with, in more recent accounts, an attempt to specify the sort of process that gives rise to them (Ariew, 1999; Samuels, 2002; Mallon & Weinberg forthcoming). The short of it is: a principle might be innate whether or not it literally figures in a computational process.

4. This is not the only way to map the external and internal projects on to levels of description. For example, one could view the two projects as distinct descriptions of what the computational specification (Level 2) requires. Nothing in the present discussion hangs on the uniqueness of my mapping.

5. When Hauser et al. write that "we find it clear that some distinctions (e.g., the agent's gender) do not carry any explanatory weight" (p. ••), they are making judgments, like Rawls's, that seem to reflect on what sort of considerations are properly considered moral ones. However, there seems little reason to think evolution respected such niceties in making us up, so it is not clear why they think such an exclusion reveals competence (computationally understood).

6. Stich (2006) has recently argued for a similar thesis albeit on distinct grounds.

3.2 Resisting the Linguistic Analogy: A Commentary on Hauser, Young, and Cushman

Jesse Prinz

In the eighteenth century, it was popular to suppose that each human capacity was underwritten by a specialized mental faculty. This view was championed by phrenologists well into the nineteenth century and then rejected by behaviorists in the early twentieth century. In contemporary cognitive science, faculties are back in vogue, due largely to the influence of Noam Chomsky's work on universal grammar. In addition to the language faculty, contemporary researchers also postulate dedicated faculties for reasoning about psychology, math, physical objects, biology, and other domains that look like a list of university departments. Conspicuously absent from this list is a faculty dedicated to morality. This was the most popular faculty of all, back in the days when men wore white wigs, and it is long overdue for a comeback. In their stimulating chapter Marc Hauser, Liane Young, and Fiery Cushman postulate an innate system dedicated to morality, and they speculate that it is interestingly similar to Chomsky's universal grammar. Related views have also been defended by Mikhail (2000), Dwyer (1999), and Rawls (1971). Hauser et al. do much to sharpen the language analogy, and they also bring recent empirical findings to testify in its defense. I applaud these contributions. Their hypothesis deserves serious attention, and their experimental findings provide data that any naturalistic theory of moral psychology must accommodate.

That said, I think it is premature to celebrate a victory for the moral faculty. There are alternative explanations of the current data. Instead of deriving from an innate moral sense, moral judgments may issue from general-purpose emotion systems and socially transmitted rules. Like art, religion, and architecture, morality might be an inevitable by-product of other capacities rather than an ennobling module. In what follows, I raise some questions about the linguistic analogy, I express some doubts about the innateness of a moral faculty, and I sketch a nonnativist interpretation of the experimental findings that Hauser et al. present. I do not take my

objections to be decisive. Hauser et al. may be right. Rather, I offer a nonnativist alternative with the hope that the dialogue between faculty theorists and their detractors will help guide research.

The Linguistic Analogy

Hauser et al. believe that there are a number of similarities between morality and language. They say that both capacities

- have an innate universal basis,
- are vulnerable to selective deficits,
- exploit combinatorial representations,
- and operate using unconscious rules.

If all four points of comparison are true, then there is indeed an analogy to be drawn between language and morality. I am skeptical about each point, but before making that case, I must enter a further point of concern. Notice capacities other than language, such as vision and motor control, are underwritten by mechanism that have each of the items on this list. Thus, the "language analogy" might equally be called the "vision analogy" or the "motor analogy." By drawing an analogy with *language* in particular, Hauser et al. are implying further points of comparison that may not hold up when all the evidence is in. Consider five potential disanalogies.

First, language has a critical period. This may be true of some perceptual systems too, but studies of, for example, vision restoration late in life suggest that language may be somewhat unusual in this respect. We don't know if there is a critical period for morality, but there are anecdotal reasons for doubt. Case studies of children who were raised in isolation, such as Genie or the wild boy of Aveyron, do not report profound moral deficits. Moreover, people can also acquire new moral values late in life, as happens with religious conversion, feminist consciousness raising, and a general trend from liberal to more conservative values that can be traced across the life span. Unlike language, learning a second morality does not seem fundamentally different than learning a first.

Second, language is usually learned in the absence of negative or corrective feedback. Is this true in the case of morality? Arguably not. Children are punished for making moral mistakes: they are reprimanded, socially ostracized, or even physically disciplined. Children also hear adults expressing negative moral attitudes toward numerous events. Of course, kids are never explicitly taught that it's worse to push people off of footbridges than to kill them by switching the course of a speeding steam

engine, but these specific rules may be extrapolated from cases acquired through explicit instruction, as I will suggest below.

Third, according to leading theories of grammar (e.g., Chomsky's government and binding theory), linguistic rules are parameterized: they have a small set of possible settings that are triggered by experience. Hauser et al. explicitly endorse this view for morality, but it's not clear what the parameters are supposed to be. Consider opposing moral systems, such as liberalism and conservatism. It doesn't look like the conflicting values are simply different settings on the same basic formation rules. Where linguistic parameter settings correspond to structural variations in how to combine primitives, variation in moral values does not seem to be structural in this sense. Consider the moralized political debate on social welfare: should governments give aid to those in need, or should the distribution of wealth be determined entirely by what individuals manage to attain in the free market? This question concerns a conflict between principles of equality and equity, rather than a conflict between alternative settings for the same basic principle. Or consider the debate about capital punishment; the two positions are dichotomous (pro or con), and they stem from different conceptions of punishment (retribution and deterrence). Similar considerations apply to debates about gender equality, gun control, and the moral permissibility of imperialism. These differences cannot be treated as parametric variations, except by trivializing that idea—that is, treating each contested policy as a parameter in its own right, which can be switched on or off. Haidt and Joseph (2004) argue that political conservatives have moral systems that contain categories of rules (e.g., rules about hierarchy, honor, and purity) that are not part of liberal morality, rather than mere variations on rules of the kind liberals share. Of course, there are some classes of rules that crop up in most moral systems, such as prohibitions against harm, but the variations in these rules are open-ended rather than parametric. Who may you harm? Depending on the culture, it can be an animal, a child, a criminal, a woman, a member of the out-group, a teenager going through a right of passage, a person who is aggressing against you, an elderly person, and so on. The range of exceptions is as varied as the range of possible social groups, and there is equal variation in the degree to which harm is tolerated (brief pain, enduring pain, mutilation, disfigurement, death). If moral rules were parameterized, there should be less variation.

Fourth, when two languages differ in grammar, there is no tendency to think one grammar is right and the other one wrong. We never start wars to snuff out people who place nouns before adjectives. In contrast,

participants in moral conflicts assume that their values are the only acceptable values.

Fifth, language uses various levels of representation: phonology, syntax, and semantics, each of which may subdivide into further levels. There doesn't seem to be an analogous range of moral levels of representation.

Of course, Hauser et al. can concede these points of contrast and restrict their analogy to the four similarities laid out above. That would weaken the language analogy, but it wouldn't undermine it. However, each of the four alleged similarities is itself subject to doubt. Let's have a look.

Do moral rules operate unconsciously? To support this claim, Hauser et al. show that people are bad at justifying their moral judgments. However, this is evidence for unconscious rules only if we think those rules should take the form of justifying principles. Suppose that moral rules take the form of simple injunctions: it's horrible to intentionally kill someone; it's pretty bad to let someone die; we have special obligations to people close to us; incest is seriously wrong; stealing is wrong too, but not as bad as physically harming; and so on. These rules are certainly accessible to consciousness. They are usually much more accessible than the rules of language.

Are moral rules combinatorial? This is a bit more complicated. As Hauser et al. point out, we certainly need a combinatorial system for categorizing actions. But notice that action categorization is something we do quite independent of morality. Our capacity to tell whether something was done intentionally, for example, operates in nonmoral contexts, and individuals who lack moral sensitivity (such as psychopaths) are not impaired in recognizing actions or attributing intentions. Psychopaths can recognize that someone is intentionally causing pain to another person. Moral rules take these combinatorial, nonmoral representations of actions as inputs and then assign moral significance to them. The distinctively moral contribution to a rule such as that killing is wrong is not the representation of the action (killing), but the attitude of wrongness. It's an interesting question whether moral concepts such as "wrong" have a combinatorial structure; they may. However, by focusing on the combinatorial structure of action representations, Hauser et al. fail to show that representations specific to the moral domain are combinatorial.

Is morality vulnerable to selective deficits? I just mentioned psychopaths, who seem to have difficulty understanding moral rules. This can be inferred from the fact that psychopaths don't exhibit moral emotions, they engage in antisocial behavior, and they fail to distinguish between moral and conventional rules (Blair, 1995). However, psychopathy is not

a selective deficit in morality. Psychopaths have other problems as well. They seem to suffer from a general flattening of affect, which also affects their ability to recognize emotional facial expressions and to recognize emotion intonation in speech (Blair, Mitchell, Richell, Kelly, Leonard, Newman, & Scott, 2002). Psychopaths may also suffer from a range of executive disorders. They tend to be disinhibited, and they make cognitive errors as a result (e.g., errors on maze tasks; Sutker, Moan, & Swanson, 1972). The moral deficit in psychopaths may result from their general emotion deficit. With diminished negative emotions, they don't experience empathy or remorse, and that leads them to be dangerously indifferent to the well-being others. If this analysis is right, then psychopathy is a domain-general problem with moral repercussions. I know of no case in the clinical literature in which morality is impaired without comorbid impairments of other kinds, most notably emotional impairments.

Is morality innate and universal? This question requires a bit more discussion.

Moral Judgments and Innateness

Elsewhere I have defended the claim that morality is not innate (Prinz, volume 1 of this collection, forthcoming-a, forthcoming-b). I will not rehearse all my arguments against nativism here, but I want to highlight some issues of contention that can help focus the debate.

To decide whether moral judgments are innate, we need a theory of what moral judgments are. Hauser et al. review several different accounts of moral judgment, or, at least, how moral judgments relate to reasoning and emotion in information processing. On one model, which I'll call "Reasons First," things proceed as follows: we perceive an event, then reason about it, then form a moral judgment, and that causes an emotion. On an Emotions First model, the sequence goes the other way around: we perceive an event, then we form an emotion that causes a moral judgment, and then we reason about it. On their view, neither of these is right. Instead, they favor an Analysis First model: we first perceive an event, and then analyze it in terms of component features such as INTENTION, AGENT, RECIPIENT, HARM; this leads to a moral judgment, which can then give rise to emotions and reasoning. I think Hauser et al. are absolutely right that moral judgment typically requires action analysis, but they are wrong to deny that other theories leave this part out. One cannot make a moral judgment about an event without first categorizing that event. Only a straw version of the Reasons First and Emotions First models would leave

out some kind of action analysis. Still, there are two important differences between the Hauser et al. model and these others. First, for Hauser et al., action analysis is not done by domain-general mechanisms that is used for categorizing actions; rather, it is done by the moral faculty, which analyzes actions using features that may be proprietary to making moral assessments. Second, for Hauser et al., both emotion and reasoning occur after moral judgments are made. So their model is a genuine alternative to these others.

Of these three models, I am most sympathetic to Emotions First, but my view pushes that approach even farther. On the Emotion First model that Hauser et al. consider, emotions *cause* moral judgments. Jonathan Haidt (2001) favors such a view, but he never tells us exactly what moral judgments are. For example, he doesn't tell us what concept is expressed by the word "wrong." Hauser et al. don't tell us the answer to that question either. I think the concept expressed by "wrong" is constituted by a sentiment. A sentiment is the categorical basis of a disposition to experience different emotions. The sentiment that constitutes the concept wrong disposes its possessor to feel emotions of disapprobation. If I judge that stealing is wrong, that judgment is constituted by the fact that I have a negative sentiment toward stealing—a sentiment that disposes me to feel angry at those who steal and guilty if I myself steal. On any given occasion in which I judge that something is wrong, I will likely experience one of these emotions, depending on whether I am the author of the misdeed or someone else is. (And likewise for other moral concepts.) Thus, in place of the Emotions First model, on which emotions cause moral judgments, I favor an Emotion Constitution model, according to which emotions constitute moral judgments. More fully elaborated, I think moral judgment involves the following sequence: first, we perceive an event and categorize it; if that event type matches one toward which we have a stored sentimental attitude, the event triggers the relevant emotion in me (e.g., guilt if it's my action and anger if it's yours). The resulting mental state is a representation of perception of an action together with a sentimental toward that action, and this complex (action representation plus emotion) constitutes the judgment that the action is wrong. The moral judgment is not a further stage in processing following on the heels of the emotion but is constituted by the emotion together with the action representation. After that, I might reason, or put my judgment into words, or reassess the case and adjust my sentiments, and so on.

I can't defend this theory of moral judgment here. The evidence is both philosophical and empirical. The empirical evidence is the same as

the evidence used to support the Emotions First model: emotions seem to occur when people make moral judgments, emotion induction alters moral judgments, and emotion deprivation (as in the case of psychopathy) leads to deficits in moral judgment. However, the Emotion Constitution model has an advantage over the Emotion First model: it is more parsimonious. Rather than saying moral concepts are mental entities that are caused by moral emotions, I say they are constituted by moral emotions. This fits with the pretheoretical intuitions. A person who feels guilty or outraged about some event can be said, *in virtue of those emotions*, to have a moral attitude about that event. This suggests that emotions constitute moral attitudes. Hauser et al. will presumably disagree. For present purposes, I simply want to explore what implications this approach to moral judgment has for nativism.

If moral judgments are constituted by emotions, then the question of whether morality is innate boils down to the question: how do we come to have the emotions we have about things such as stealing, killing, cheating, and so on? A nativist will propose that we are innately disposed to have these emotions in virtue of domain-specific principles (which may be parameterized). Here's a nonnativist alternative. Suppose that a child who has no moral attitudes or moral faculty engages in a form of behavior that her caregivers dislike. The caregivers may get angry at her, and they may punish her in some way. For example, they might scold her or withdraw love and affection. Children rely on the affection of caregivers, and when punished, those all-important attachments are threatened. The emotion elicited by threats to attachment is sadness. Thus, a child who misbehaves will be led to feel bad. Over time, she will associate that feeling of sadness with the action itself; she will anticipate sadness when she considers acting that way again. Once the child associates sadness with that action, we can say she feels regret, remorse, or even guilt about it. These moral emotions can be defined as species of sadness directed at courses of action. The main difference between ordinary sadness and guilt is that guilt promotes reparative behavior. Such behaviors need not be innate. They are a natural coping strategy for dealing with cases where you have angered another person. The child who is punished will also come to have the same anger dispositions as those that punish her. Children are imitative learners. If a child sees her parents get angry about something that she does, she will feel sad about it, but she will also come to feel angry at other people when they engage in that behavior. She will copy her caregiver's reactions. This will also allow children to acquire moral rules concerning behaviors that they have never attempted, such as prohibitions against murder and rape.

When such behaviors are mentioned by caregiviers, there is almost always an expression of emotion. When we mention things that we morally oppose, we do not conceal our emotions. Children imitatively pick up these attitudes. Notice that this story explains the transmission of moral rules by appeal to domain-general resources: children must be able to categorize actions, they must experience sadness when punished, and they must be disposed to imitate anger and other negative emotions expressed by caregivers. If a child has these capacities, she will learn to moralize. She does not need an innate moral sense.

This developmental just-so story is intended as a possible explanation of how one could learn moral rules without having an innate moral faculty. If moral judgments are sentimental, then moral rules are learnable. However, it is one thing to say that moral rules are learnable and another thing to say they are learned. After all, we could be born with innate moral sentiments or sentimental dispositions. Just as we are biologically prepared to fear spiders, we might be biologically prepared to feel angry and guilty about various courses of action. We need a way of deciding whether moral rules are innate or acquired. One way to approach this question is development. Do children acquire certain moral rules more easily? Are others impossible to acquire? Are certain moral rules learned without punishment, or other kinds of social interaction that condition emotional responses? I think these are all important open questions for research. I do think that there is extensive evidence for the claim that punishment plays a central role in moral education (Hoffman, 1983), and that leads me to think that moral nativism will be difficult to defend by appeal to a poverty-of-the-stimulus argument, as I mentioned above. I also think that the wide range of moral rules found cross-culturally suggests that children can acquire moral attitudes toward just about anything. However, both of these observations are anecdotal, and it is crucial at this stage to systematically search for innately prepared moral rules.

Trolley Cases

In suggesting that morality may not be innate, I don't want to deny that we are innately disposed to engage in some forms of behavior that are morally praiseworthy. Perhaps helping behavior, reciprocal altruism, and various forms of peacemaking are species typical in the hominid line. But there is a difference between behaving morally and making moral judgments. My hypothesis is that people are not innately equipped with a faculty of moral judgment. Moral concepts, such as right and wrong, are

acquired from domain-general mechanisms. The fact that we are innately disposed to do some praiseworthy things is no more evidence for innateness of a moral sense than is the fact that we are disposed to take care of our young. Laudable behavior can exist without the capacity to praise it as such. One of the exciting features of Hauser et al.'s research program is that they are directly investigating moral judgments, rather than morally praiseworthy behavior. Their research on trolley cases can be interpreted as an argument for innate moral judgments.

Here's how I interpret that argument. There are moral judgments about moral dilemmas that are very widespread, homogeneous across different demographics, and demonstrable across cultures. These judgments do not seem to be learned through explicit instruction, and they do not seem to be based on consciously accessible reasoning processes. Together, this pattern is consistent with the conclusion that the judgments issue from an innate moral faculty. It's not a demonstrative argument, of course, but it's a reasonable argument to the best explanation—or at least it would be, if there weren't other equally good nonnativist explanations available.

Here's how a nonnativist might account for the data. On my view, there is a moral rule of the form "Intentionally taking another person's life is wrong." This rule consists of a domain-general action representation (intentionally taking a person's life) and a sentiment (which disposes one to feel angry or guilty if a person is killed by someone else or by oneself). The nonnativist needs to explain how such a rule could come about without being hardwired. That does not look like an insuperable challenge. Societies that allow killing, at least within the in-group, are not very stable. In very small-scale societies, built around extended kin groups, there may not be a need for any explicit rule against killing. We rarely have motives to kill our near and dear, especially if we feel a sense of attachment to them. However, as societies expand to include many nonrelatives, pressure arises to introduce killing norms, and that probably doesn't take much work. If you try and kill your neighbor, he and his loved ones will get pretty miffed. Other members of the community, afraid for their own security, may get upset too, and they will try punish you or banish you. Thus, aggression against others naturally elicits strong reactions, and those reactions condition the emotions of the aggressor. Knowing that aggression can lead to alienation and reprisal, you resist. When you think about aggressing, you feel anticipatory guilt, and, when you imagine others aggressing, you get angry about the harm they will do. Thus, we don't need innate strictures against killing, because the natural nonmoral emotions that are elicited by acts of aggression will instill the sentiments that

constitute moral disapprobation. The rules against killing may, at first, be limited to the in-group, because aggression against more distant strangers may go unpunished by the local community. However, when communities become more transient, more diverse, or more dependent on others for trade, strictures against killing generalize, because harming distant strangers can be perceived as a potential threat to members of the local group.

So much for the genealogy of norms against killing. The nonnativist also needs to explain helping norms. Most of us think we should help people in need if we can do so at little personal cost. Is this an innate rule? Not necessarily. It could easily emerge through cultural evolution, because helping confers obvious advantages. If I join a group whose members will help me when I am in need, I will fare better than if I join a group of selfish people. However, helping always introduces free-rider problems. How can I be sure that people in my community will help me? Game theoretic models suggests that the best solution for coping with free riders is punishment. If I penalize people for being unhelpful, then they will be more likely to help in the future. Punishment leads people to feel guilty about free riding and angry at other free riders. Thus, when unhelpful individuals are punished, emotions are conditioned, and a moral attitude in born. In sum, I think the social and emotional consequences essentially guarantee that most societies will end up with moral rules about killing and helping. Nonviolent cooperation may be a precondition to stability in large populations. However, these rules about killing and rules about helping may differ from each other in one respect. Several factors are likely to make killing norms stronger than helping norms. First, in cultural evolution, prohibitions against killing are more vital than prohibitions against unhelpful behavior, because a group whose members kill each other will fare worse than a group of members who go out of their way to help each other. Second, helping also carries more personal cost than refraining from killing. Third, acts of aggression naturally elicit fear and anger, so it is easier to inculcate strong sentiments toward killing. Collectively, these factors essentially guarantee that sentiments toward killing will be stronger than sentiments pertaining to helpful and unhelpful behavior. If the Emotion Constitution model of moral judgment is right, this difference in sentimental intensity is tantamount to a difference in the strength of the respective moral rules.

I have been arguing that we can account for norms about helping and killing without supposing that they are innate. Once they are in place, they can guide behavior, and, on occasion, they will come into conflict. When

this happens, there are two factors that will determine which rule will win. One factor is the extent to which actions in the event under consideration can be construed as instances of killing, on the one hand, or helping, on the other. Failure to conform to paradigm cases of either will diminish the likelihood that we will apply our rules about killing and helping. If some course of action is only a borderline case of killing, we may apply our killing rule with less force or confidence. For example, suppose someone causes a death as a side effect of some other action. This is not a paradigm case of killing. In terms of cultural evolution, groups have greater interest in condemning people who form direct intentions to kill than people who kill as a side effect, because the person who will kill intentionally poses a greater threat. Killing without the explicit intention to kill is a borderline case of the rule. The other factor is emotional intensity. For example, if we can help a huge number of people, our helping rule may become emotionally intense. In some cases, emotions may be affected by salience: if attention is drawn to an act of helping or killing, the corresponding rule will be primed more actively, and the emotions will be felt more strongly.

Now at last, we can turn to the trolley cases presented by Hauser et al. These cases are interesting because they pit helping norms against killing norms. We can now see whether the nonnativist, emotion-based theory can explain the results. In the first case, Frank is on top of a footbridge and can push a man into the path of a trolley, thereby saving five people further down on the track. Only 11% of subjects think it's okay to push the man. One explanation is that this is a paradigm case of killing, and the killing rule is, all else being equal, more emotionally intense than the helping rule. It's also a very salient case of killing, because subjects have to imagine Frank pushing someone, and the thought of physical violence attracts attention and increases emotion. In a second case, Denise can pull a lever that will guide a trolley down an alternate track, killing one person, rather than allowing it to kill the five people on the track it is currently on. Here 89% say it's permissible to pull the lever. The numbers change because this is not a paradigm or emotionally intense case of killing. The person who is killed is not physically assaulted, and Denise does not need to form the intention "I want to cause that guy's death."

The next case is a bit puzzling at first. Like Denise, Ned can pull a lever that will send a train on a different track, killing one rather than five. However, unlike the Denise case, in Ned's case the track is a loop that would reconnect with the original track and kill the five people were it not for the fact that the guy on the alternate track is heavy enough to stop the trolley in its tracks. In this situation, only 55% of subjects think Ned is

permitted to pull the lever, killing one and saving five. Why would the minor addition of a looping track change permissibility judgments from the Denise case? The answer way be salience. When we imagine a person being used to stop a trolley in its path, the imagery is more violent and more emotionally intense. It is also a more paradigmatic case of killing, because Ned has to explicitly form the intention that the person be crushed; otherwise, the train wouldn't stop.

Hauser et al.'s final case is a slight variant on the Ned case. Here, Oscar can pull a lever that will send a train on a loop track that is obstructed by a large weight; the weight will prevent the train from rejoining the original track where it would kill five, but, unfortunately, there is a man standing in front of the weight who will be killed if the lever is pulled. Seventy-two percent of subjects think this is permissible. These permissibility ratings are higher than in the Ned case, because it is a less paradigmatic case of killing: the death in the Oscar case is an accidental by-product of sending the train into the weight. There is just one remaining question: why are the permissibility ratings in the Oscar case slightly lower than in the Denise case? The answer may involve salience. In the vignettes, the solitary man in the Oscar case is introduced with a 20-word sentence, and the solitary man in the Denise case is introduced with 10 words. In the Oscar case, that man is crushed between the train and the weight, and in the Denise case, he is killed the same way that the five people on the other track would have been killed. Thus, the Oscar case draws extra attention to the victim. These explanations are sketchy and tentative. I offer them to illustrate a simple point. If one can tell a nonnativist and sentimentalist story about moral rules pertaining to killing and helping, there are resources to explain intuitions about trolley cases. Without ruling out this alternative account, Hauser et al.'s argument for nativism loses its force. At this stage, it's fair to say that both the nativist and the nonnativist accounts are in embryonic stages of development, and both should be considered live options as we investigate the origin of our capacity to make moral judgments.

The account that I have been proposing leads to some predictions. The first is consistent with Hauser et al.'s account, the second is slightly harder for them to accommodate, and the third is is more naturally predicted by my account. First, I think that moral rules contain representations of actions, and these representations may take the form of prototypes or exemplars (e.g., a typical murder). I predict that the moral judgments will weaken as we move away from these prototypes. Hauser et al. may agree.

Second, I think that helping and harm norms are socially constructed to achieve stability within large groups, and consequently, there may be subtle cultural differences as a function of cultural variables. For example, consequentialist thinking may increase for groups that are highly collectivist (hence more focused on what's best for the collective), for groups that are engaged in frequent warfare (hence more desensitized to killing), and for groups that are extremely peaceful (where norms against killing have never needed to be heavily enforced). In highly individualist societies, there is less overt focus on helping behavior, and consequentialist thinking may diminish. Likewise, in highly pluralistic societies, pluralism promotes the construction of strong rules against killing, because such rules are often needed to ensure peace in diverse groups. Hauser et al. report on some cross-cultural work, but there are two limitations of the data they report. First, as they note, their non-American subjects understand English and have access to computers, so they are probably similar to us. Second, Hauser et al. do not report the actual percentages for their cross-cultural samples; so even if every tested culture tended to say Frank's behavior is less permissible than Denise's, the actual percentages who hold that dominant view may differ. It is important to note that Hauser et al. *can* allow variation in moral judgments. The language analogy predicts that principles will have parameters that get set differently in different contexts. My worry is that this is the wrong kind of variation. In language, switching parameters results in differences that are qualitative and arbitrary. The differences that I am imagining are quantitative and tailored to fit cultural variables. That is suggestive of learning rather than innateness.

Third, the Emotion Constitution model predicts that manipulation of emotions should influence judgments on trolley dilemmas. By making one of the two courses of action more salient, more violent, more personal, or more emotionally evocative in some other way, one should be able to alter the permissibility ratings. Psychopaths should not be influenced to the same degree by emotional manipulations. Such findings would count against Hauser et al.'s nonaffective theory of moral judgment, and they would also count against the view that moral judgments are driven by domain-specific (or at least encapsulated) mechanisms.

If these predictions pan out, they add support the emotion-constitution model. That model is compatible with nativism, but it also lends itself to a plausible nonnativist account of how we come to acquire moral rules. In this commentary, I haven't provided strong evidence for the nonnativist view or against the view favored by Hauser et al. Rather, my goal has been

to suggest that, at this early stage of inquiry, several models remain compatible with the evidence. Hauser et al. would undoubtedly agree, and, in the coming years, we will need to find ways to test between these options. Let me sum up with a few questions for Hauser et al. that highlight places where their model and my alternative come apart. Why think that the analyses of action that precede moral judgment are carried out by a domain-specific moral faculty? Why think that emotions arise as consequences of moral judgments rather than causes or constituent parts? Why think that moral principles are innate rather than learned solutions to problems facing all cultures? And what is it about language, as opposed to any other faculty, that sheds light on our moral capacities? Hauser et al. have embarked on an important research program, and the linguistic analogy has been a valuable source of inspiration. My hunch is that it will eventually prove more productive to drop that analogy and adopt a model that places greater emphasis on learning. For now, we can make most progress by keeping both approaches on the table.

3.3 On Misreading the Linguistic Analogy: Response to Jesse Prinz and Ron Mallon

Marc D. Hauser, Liane Young, and Fiery Cushman

Oscar Wilde noted "Always forgive your enemies—nothing annoys them so much." Before we forgive our critics, however, we thank Prinz and Mallon for their thoughtful comments, and for taking the linguistic analogy as a serious proposal amid the current excitement at the interface between moral philosophy and moral psychology. What we forgive is their targeted comments on several issues that are either irrelevant to the linguistic analogy or premature given that we know so little about the nature of our moral psychology. Some of the confusion is undoubtedly due to our own exposition, and some to the rapid pace of theoretical and empirical developments that have emerged since we submitted the final draft and received the commentary.

We begin by clarifying the main goals of the linguistic analogy, including, most importantly, its unique set of empirically tractable questions and challenges. Our hope is that this response, guided by Prinz and Mallon's comments, serves as the next installment on a much larger project that, we can all agree, will yield interesting results irrespective of the strength of the analogy. The reason for this is simple: until the questions that emerge from the analogy are taken seriously, and pitted against the alternatives, we will have only a weak understanding of the mature state of moral knowledge, how it is acquired within the individual and species, and the extent to which it relies upon domain-specific machinery. In this sense, we see the arguments generated in our target essay, and developed more fully elsewhere (Dwyer, 2004; Hauser, 2006; Mikhail, 2000, in press), as analogous to the minimalist program in linguistics (Chomsky, 1995, 2000): a set of fascinating questions with ample room for movement on theoretical, empirical, and methodological fronts.

For a novel research program to breathe, it is important that its claims be properly understood and that challenges be targeted at the proper level. Let us start then by highlighting two important points of agreement: both

Prinz and Mallon (1) endorse our research program focused on the cognitive systems responsible for generating the basic representations that serve as input to the process of moral judgment and (2) support our position that these systems operate over the representations of actions, intentions, causes, and consequences. By supporting these two points, they at least implicitly support a third which, we submit, follows: some moral principles are formulated over the core representations that enter into our moral judgments. The primary thrust of the linguistic analogy is to study these systems and bring them to the attention of philosophers and psychologists. It is in this spirit that we turn next to a more detailed look at the linguistic analogy, pinpointing what we perceive as its central assumptions and predictions, together with a body of relevant data. Along the way, we point out some of the challenges raised by Prinz and Mallon, including the nonnativist alternative based on emotions and real-world experiences, and emphasize the need to posit an innate, dedicated moral organ.

Both Prinz and Mallon attribute to us the view that the cognitive systems responsible for generating basic representations used in moral judgment are in fact specific to the domain of morality. This is not our view—indeed, it should have been clear that we hold the opposite position. Moral judgment depends on a wide range of representational inputs generated by cognitive systems adapted for and typically engaged in entirely different functions. Analogous cognitive mechanisms support linguistic competence without being specific to the domain of language. To clarify, take the rather simple phenomenon of speech perception. Although the last fifty years of research has largely assumed that we are endowed with a dedicated neural system for processing speech, neuroimaging studies with normal subjects, together with comparative and developmental studies of other animals and infants, suggest that much of speech perception may derive from very general and ancient auditory mechanisms. For example, a recent study by Vouloumanos, Hauser, and Werker (unpublished manuscript) showed that neonates less than 48 hours old evidenced no preference for human speech over rhesus monkey vocalizations. Similarly, comparative studies of human adults, infants, and cotton-top tamarin monkeys revealed no difference in the capacity to use transitional probabilities to segment a continuous stream of speech (Hauser, Newport, & Aslin, 2001). These results suggest that early stages of speech perception and segmentation are not mediated by processes that are specific to the domain of language.

Though we explicitly recognize the role of domain-general mechanisms, we are nonetheless committed to the existence of some cognitive mechanisms that are specific to the domain of morality. These we term the "moral faculty." These systems are not responsible for generating representations

of actions, intentions, causes, and outcomes; rather, they are responsible for combining these representations in a productive fashion, ultimately generating a moral judgment. Our thesis is that the moral faculty applies general principles to specific examples, implementing an appropriate set of representations. We refer to these principles as an individual's "knowledge of morality" and, by analogy to language, posit that these principles are both unconsciously operative and inaccessible.

Mallon notes that we must distinguish between a theory that can adequately account for the pattern of people's moral judgments and a theory that is actually instantiated in people's heads. We fully agree, especially since this captures the parallel distinction in linguistics. To be precise, we must distinguish between a set of principles that are descriptively consistent with people's moral judgments and the principles that people in fact carry around in their heads, doing the work of adjudicating between moral rights and wrongs. As Mallon correctly intuits, we are aiming at principles in the head. But the first step, of course, is to determine the set of principles at the descriptive level.

Consider the following example as an illustration of how first to identify the set of descriptive principles that are operative in guiding moral judgment and then to investigate the extent to which these principles are expressed in the course of justification. In a recent paper (Cushman, Young, & Hauser, 2006) focused on the relationship between operative and expressed principles, we develop the argument that a three-pronged approach is necessary to assess whether particular principles mediate our moral judgments and whether these principles serve as the basis for our justifications. *Prong 1*: Develop a battery of paired dilemmas that isolate psychologically meaningful and morally relevant, principled distinctions. *Prong 2*: Determine whether these targeted principles guide subjects' moral judgments. *Prong 3*: Determine whether subjects invoke these principles when justifying their moral judgments. With this approach, we explored three principles:

Action principle Harm caused by action is morally worse than equivalent harm caused by omission.
Intention principle Harm intended as the means to a goal is morally worse than equivalent harm foreseen as the side effect of a goal.
Contact principle Harm involving physical contact with the victim is morally worse than equivalent harm involving no physical contact.

Based on a sample of approximately 300 subjects, largely from English-speaking, Western countries, analyses revealed support for the three targeted principles in 17 out of 18 paired dilemmas. That is, subjects judged

harm caused by action as worse than omission, intended harm as worse than foreseen harm, and harm involving contact as worse than with no contact. When we turned to justifications, 80% of subjects recovered the key distinction for the action–omission cases, 60% for the contact–no contact cases, and only 30% for the intended–foreseen cases. This pattern suggests that the intended–foreseen distinction is operative but results in an intuitive judgment. The other principles are also operative but appear to be at least accessible to conscious awareness, to some extent.

Are the descriptive principles targeted in this study isomorphic to the domain-specific principles that constitute an individual's moral knowledge? At present we cannot say. We know that these principles are descriptively adequate to capture the observed pattern of subjects' moral judgments, but it remains a viable possibility that they exert their influence during the generation of the relevant representations that are external to and feed into moral judgment. Of course, a direct implication of the view that these principles are not specific to morality is that they influence judgments and behaviors outside the moral domain. Identifying nonmoral analogues of these descriptive principles—if indeed they exist—is an important area for future research.

Thinking about the moral faculty from this perspective leads us directly into Mallon's point that evolution may have created particular biases that set initial conditions on the valenced responses. Consider sex, and the extent to which degrees of genetic relatedness matter. An agent INTENDS/DESIRES to ±SEXUAL INTERCOURSE with X_r, where X is some sexual partner and r is his or her degree of genetic relatedness to the agent. If we ask whether sexual intercourse is morally permissible with X, the answer depends on r. Evolution appears to have set up a bias, in the sense that r values between .125 and .5 are generally coded as –SEXUAL INTERCOURSE—that is, forbidden. This may be the default setting or bias, open to modification (to some extent) by the local culture. Again, the initial valence settings may have been established on the basis of their statistical effects (e.g., the probability that mating with parents and siblings will reduce fitness) and only later hooked into the emotions as reinforcing agents. In sum, we completely agree with Mallon that evolution has set us up with strong biases. These biases may enter into moral judgments, and at this point, we are agnostic on whether they figure into moral competence or performance.

To summarize thus far, we propose, and Prinz and Mallon agree, that a deeper understanding of the sources of our moral judgments requires further research into the nature of our representations of actions, inten-

tions, causes, and consequences. The system involved in generating such representations is not specific to the moral domain. In parallel to language, however, individuals possess knowledge of morality that is comprised of domain-specific moral principles operating over these representations. Though we are only at the earliest stages of this research program, our empirical studies suggest a methodology to determine candidate principles for domain-specific moral knowledge. Whether the descriptive principles that capture patterns of moral judgment in fact characterize features of the moral faculty or features of the cognitive systems that feed into the moral faculty is presently unknown, but, we submit, not unknowable.

What we wish to stress is that the linguistic analogy provides a substantive foundation for constructing testable hypotheses and collecting the relevant data. For example, as a theory, it demands a proper descriptive account of the mature state of moral knowledge. Until we understand our moral psychology at this descriptive level, including some subset of its principles, it is virtually impossible to make progress on other fronts, including, especially, issues of moral acquisition (explanatory adequacy in Chomsky's terms), domain-specificity, characteristic neural breakdown, and evolutionary origins. That is, we need to understand the nature of our mature subject's moral knowledge before we can ask how it evolved, develops, and is instantiated in neural tissue.

A thorough characterization of moral knowledge is particularly critical to adjudicate between nativist and empiricist claims. For example, Prinz states that he doubts there is a critical period for morality in the same way that there is for language or that learning a second moral system is like learning a second language. However, we are only able to determine that there is a critical period for language because we have a relatively deep understanding of the principles underlying the mature state of linguistic knowledge and, thus, can see what happens to the externalization of such knowledge in expressed language as a function of severe developmental isolation. Furthermore, we are only able to contrast native and second language acquisition because we understand *what* is being acquired. On the basis of a clearly characterized linguistic target, often articulated in terms of principles and parameters, we can state that native language acquisition is fast, effortless, untutored, and relatively immune to negative evidence or correction. Second-language acquisition is slow, effortful, tutored, vulnerable to negative evidence and correction. Surprisingly, no one has ever systematically compared the acquisition of native and second moral systems.

We end here with a discussion of the role of emotions in guiding our moral psychology and behavior. Though many of the questions that emerge from adopting the linguistic analogy have little or nothing to do with the emotions, our perspective puts into play a different way of looking at the role of emotions. To clarify, consider three ways in which emotions might enter into our moral judgments. First, an individual's emotional response to a particular circumstance might influence the representations he forms of the actions, intentions, causes, and consequences associated with that circumstance. Second, an individual's emotional response to a particular circumstance might, itself, be among the representational inputs to the moral faculty. This characterization implies the existence of a domain-specific moral principle such as "If it produces negative affect, it is morally wrong." Finally, it is possible that emotion has no influence upon moral judgment but is only a product of it.

Prinz proposes "the emotion-constitution model, according to which emotions constitute moral judgments" (p. ••). This corresponds most closely to our second possibility, but with some potential differences. On the one hand is the rather trivial and uncontroversial claim that moral judgments are not synonymous with negative emotion. There are many instances in which we experience a negative emotion in the absence of moral disapproval (e.g., anger from stubbing a toe, disgust from seeing blood). On the other hand, Prinz appears to define moral judgment as a variety of negative emotion, such that the meaning of wrong is the feeling of wrongness. Stranding the problem here simply raises another: how does one determine wrongness in the first place? Prinz's solution is that "the concept expressed by 'wrong' is constituted by a sentiment . . . [which is] the categorical basis of a disposition to experience different emotions" (p. ••). In essence, Prinz is describing a mechanism that has at its disposal some categorical basis (principles) that presumably operates over some set of representations and that outputs emotions that we label as "right" or "wrong" (moral judgments). Ironically, then, what Prinz calls a "sentiment" is apparently identical to what we call the "moral faculty."

What the discussion above boils down to is that for both our perspective and the one Prinz favors, we are left with a binary choice: either emotion plays a role in moral judgments by shaping the representational input into the judgment mechanism (Prinz's sentiment, our moral faculty), or it is merely a consequence of that mechanism. This is an open and empirically tractable question that we have begun to explore. Let us illustrate with some recent patient data, acquired since our original submission, and only briefly discussed.

In collaboration with Michael Koenigs, Daniel Tranel, Ralph Adolphs, and Antonio Damasio (2007) we have explored the nature of moral judgments in six individuals with adult-onset bilateral damage to ventromedial prefrontal cortex (VMPC), an area noted for its critical role in linking emotion to decision making (Bechara, Damasio, Tranel, & Damasio, 1997). VMPC damage is associated with diminished autonomic and subjective response to passive viewing of emotionally charged pictures (Blair & Cipolotti, 2000; Damasio, Tranel, & Damasio, 1990), recall of emotional memories (Tranel, Bechara, Damasio, & Damasio, 1998), contemplation of risky choices (Bechara et al., 1997), and consideration of counterfactual outcomes (e.g., regret; Camille, Coricelli, Sallet, Pradat-Diehl, Duhamel, & Sirigu, 2004). We found that VMPC subjects were more likely to endorse personal or emotionally salient moral violations presented in hypothetical scenarios developed by Greene and colleagues (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) than were comparison groups, including normal subjects and brain damaged controls. More specifically, VMPC subjects were more likely to endorse violations that maximized aggregate welfare (e.g., throw a man off a bridge to save five others), resulting in heavily consequentialist judgments. There was no difference between VMPC subjects and comparison groups on either nonmoral or impersonal moral scenarios, showing that many aspects of their decision-making systems are intact and, significantly, that a variety of moral dilemmas can be evaluated in the absence of emotional input. A supplementary analysis of the personal moral scenarios showed that the difference between VMPC participants and comparison groups was restricted to the "difficult" as opposed to "easy" scenarios, as measured by uniformity of judgment within the comparison groups, showing further that even some judgments of emotional moral actions are intact. These analyses suggest that the effect of VMPC damage on moral judgment is both specific to its role in emotion processing and specific to scenarios for which there are no explicit adjudicating norms, that is, scenarios posing "difficult" moral dilemmas. In short, it appears that there may be an important role for emotion in shaping the representational inputs into the moral faculty under highly selective situations.

These data bear on Prinz and Mallon's concern about the notion of a moral organ. Their own view is that current work in neuropsychology does not support the idea of a dedicated, domain-specific moral organ and, if anything, supports the alternative, domain-general view. Although the existing data may be revealing with respect to moral cognition, they don't

yet illuminate the linguistic analogy. Consider the existing work on psychopaths and patients with VMPC damage. Neither group shows selective damage in the moral sphere, which Mallon and Prinz take to be strong evidence against a dedicated moral faculty. However, for both theoretical and methodological reasons, we disagree. Many of the current tests of patients thought to have deficits in the moral sphere have not addressed the issues raised by the linguistic analogy. For example, the published work on prefrontal lobe patients is based on moral reasoning tasks, in particular, Kohlberg's battery of tests, which measure moral maturity based on the content of justifications rather than the nature of the judgments. Because of their emphasis on conscious reasoning, these measures aren't particularly revealing with respect to intuitive judgments, such as those tapped by the dilemmas featured in our Web-based experiments, recent functional neuroimaging studies (Greene et al., 2004), and the new collaborative work reviewed above on moral judgment in individuals with VMPC damage (Koeniss et al., 2007). Further, all of the tests administered to psychopaths that are morally relevant focus on the conventional–moral distinction, in which subjects distinguish between unambiguous conventional transgressions and unambiguous moral transgressions, but never between right and wrong. Furthermore, such tests have not included moral dilemmas where there are no obvious norms to adjudicate between different choices, where both choices lead to harm, for example.

At a theoretical level, we are open to the possibility that even the domain-specific components of the moral faculty may be divisible into discrete units. Indeed, some of the evidence we have presented in this discussion point to just such a multisystem model. Some moral principles appear to be available to conscious reflection, while others do not. Patients with emotional deficits show abnormal moral judgments on some dilemmas, but not others. We argue that such evidence, far from delivering a blow to the linguistic analogy, is in fact an encouraging sign of the type of refinements to models of moral judgment that have been occurring for decades in the research on language. The language faculty includes subsystems for phonology, morphology, semantics, and syntax, and even these subsystems can be further divided. For example, recent work on dysgraphic patients (Miceli, Capasso, Banvegnu, & Caramazza, 2004) has revealed individuals with deficits in the representation of vowels, others for consonants, highlighting the distinctive neural foundations for these linguistically specific distinctions.

Let us end as we started with a comment by Oscar Wilde: "I choose my friends for their good looks, my acquaintances for their good characters,

and my enemies for their good intellects." We couldn't be more pleased to have such excellent "enemies" as Prinz and Mallon in an area of research that is fueled with excitement, passion, and hope for fundamental discoveries about the nature of moral thought and action. As we have tried to clarify, by drawing on analogy to language, we raise new questions about the nature of our moral psychology. In particular, we force empirically minded researchers interested in the nature of our moral judgments to tackle five distinctive questions: (1) What are the principles that characterize the mature state of moral competence? (2) How is this moral knowledge acquired? (3) How does our moral competence interface with those systems entailed in performance? (4) How did our moral competence evolve? (5) To what extent are the mechanisms underlying our moral competence domain-specific? We are nowhere near any resolution on any of these questions, and thus nowhere near a thumbs up or down for the linguistic analogy. With these questions in mind, however, and with answers forthcoming, we can be confident that our understanding of moral knowledge will rapidly deepen.