# Reviving Rawls' Linguistic Analogy: Operative principles and the causal structure of moral actions

**Article** · January 2006

CITATIONS
87

READS
1,543

**3 authors**, including:

Marc D Hauser
Risk-Eraser, LLC
**355** PUBLICATIONS **28,169** CITATIONS

SEE PROFILE

# Reviving Rawls' Linguistic Analogy:

**Operative principles and the causal structure of moral actions**

Marc Hauser[1,2,3], Liane Young[1] and Fiery Cushman[3]

Departments of [1]Psychology, [2]Organismic & Evolutionary Biology, and [3]Biological Anthropology
Harvard University, Cambridge, MA, 02138

**[in press]** *Moral Psychology and Biology, Ed. W. Sinnott-Armstrong, Oxford U. Press, NY*

**Table of Contents**



## *1.1 Introduction*

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 (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 prep), 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 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 develop others elsewhere (Hauser, in press). But 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.

## *1.2. 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 an instinct to acquire a natural language (Pinker, 1994). More narrowly, it can be thought of as the set 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. But 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 (S. R. Anderson & Lightfoot, 2000; Fitch, Hauser, & Chomsky, in press; 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 principles 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 school marm'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, in press; J. Mikhail, 2000).

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. But 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 non-speech. 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 monolog or external dialog. This stage of cognitive processing is common to both spoken and sign language. The hierarchical structure of language, together with its recursive operations and interfaces to phonology and semantics appear to be

unique to 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 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 their 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 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 (S. R. 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. But 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.

*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 its 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 6-7 million years ago. But 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 neuroimaging 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.

## *1.3. 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 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.

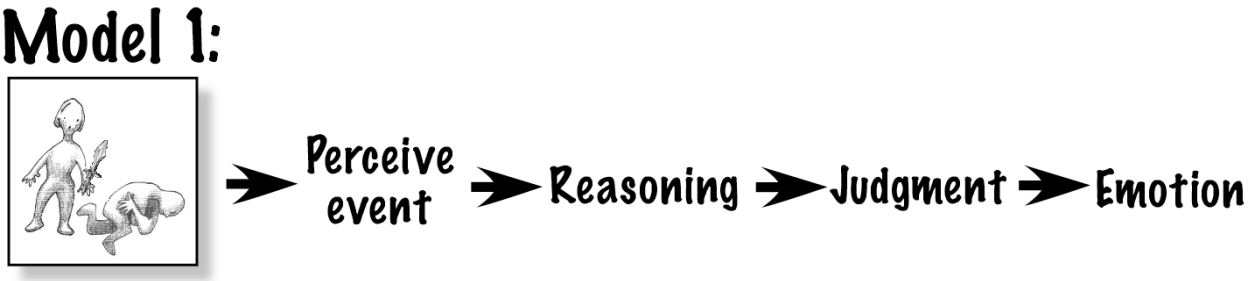


Figure 1. The Kantian creature and the deliberate reasoning model

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

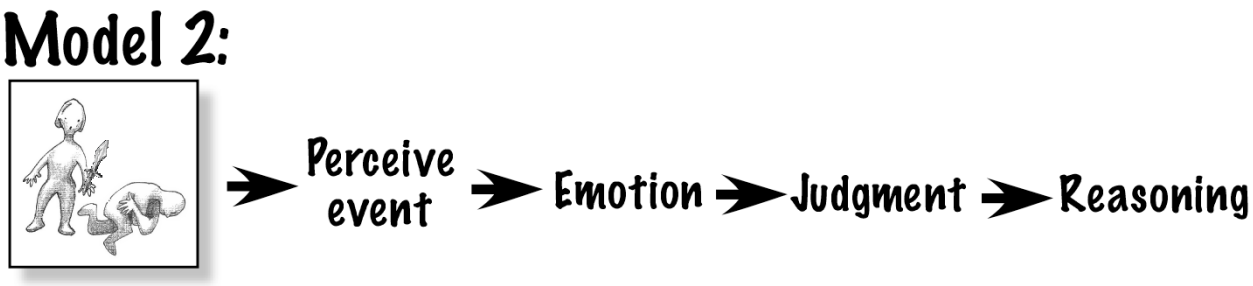


Figure 2. The Humean creature and the emotional model.

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]; 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 [see 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, 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.

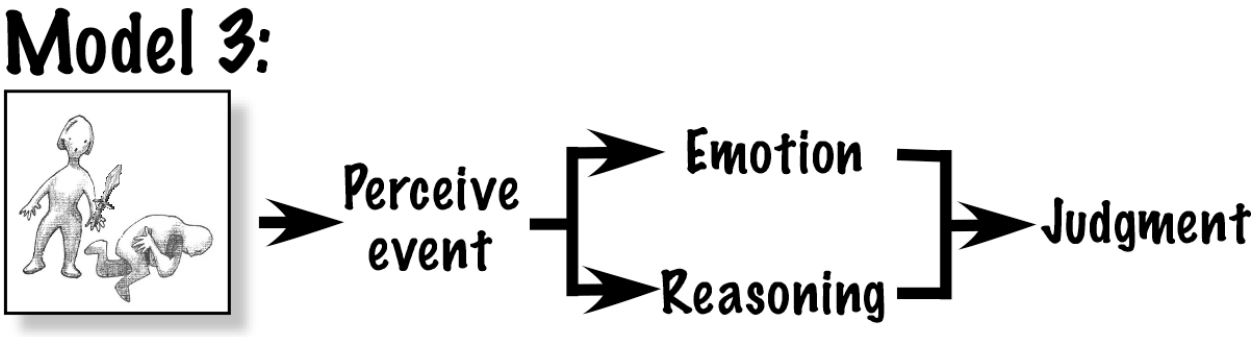


Figure 3. A mixture of the Kantian and Humean creatures, blending the reasoning and emotional models.

Independently of which account turns out to be correct, this break-down 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 causes and consequences for self and other. As Mikhail (2000; in prep), Jackendoff (2005) and Hauser [in press] 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' 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." 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' 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 4] (Hauser, in press); 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 causes and consequences of actions.

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.

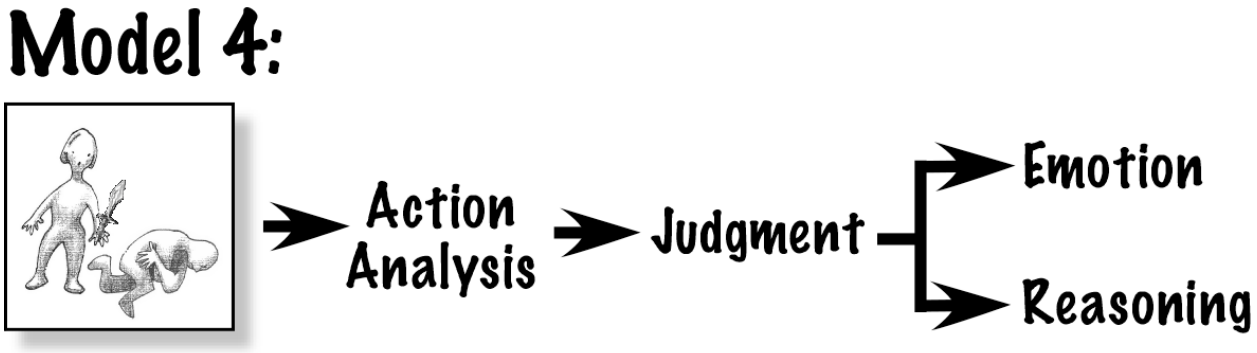


Figure 4. The Rawlsian creature and action analysis model.

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.

*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 sub-goals 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 [Figure 5]:

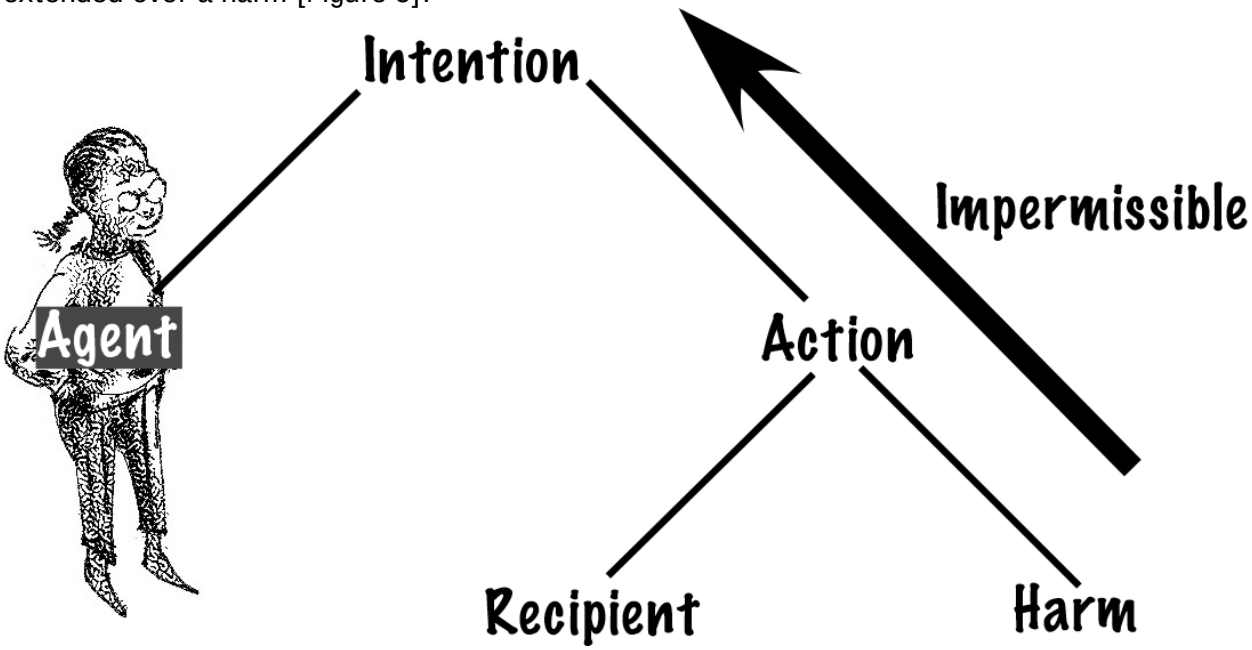


Figure 5. Some components of the causes and consequences of morally relevant actions.

In reality, the principle will be far more complicated and abstract, and include other parameters. [See Mikhail [2000; in prep] 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 mean. 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 *3-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 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 (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, independently 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 down stream, 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 and chapters in this volume by Kiehl and Kennett] we would leave our moral competence intact [i.e., moral judgments would be normal], but 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. But 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 both the problems of 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 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 (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 7-15 ounces, 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 on-line 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 (in press). We turn next to some of the empirical evidence, much of which is preliminary.

## *2.1 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, in press; Macnamara, 1990; J. 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.

## *2.2. 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, 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 causal 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 trolley-esque 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., 2004; Greene et al., 2001; 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, in press; J. Mikhail, 2000; Mikhail, in prep; 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 www.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, in prep]. 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 unambiguous benefits to the patients. The four trolley problems are presented below and illustrated in Figure 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 answering these questions, 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 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 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, 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 intutions 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?

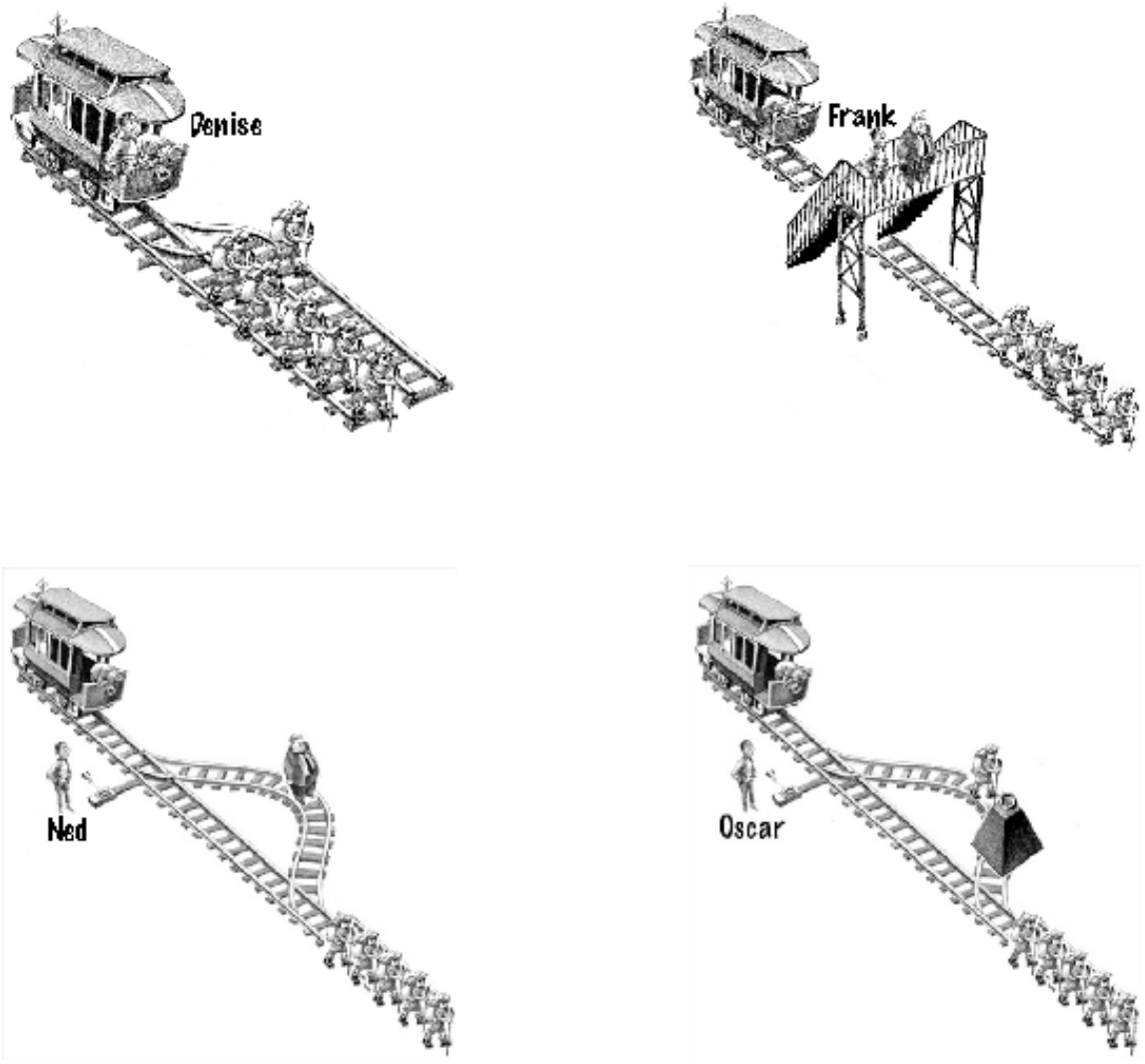


Figure 6. The core family of trolley dilemmas used in internet studies of moral judgments and justifications.

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 which were from English-speaking countries (Hauser, Cushman, Young, Jin, & Mikhail, in review). 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 1]. In the case of Ireland the effect was marginally significant at $p=.07$ with a sample size of 16 subjects.

**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 | |
| Israel | | **Ethnicity** |
| The Netherlands | **Age** | American Indian |
| New Zealand | 10-19yrs | Asian |
| Philippines | 20-29 | Black non-Hispanic |
| Singapore | 30-39 | Hispanic |
| South Africa | 40-49 | White non-Hispanic |
| Spain | 50-59 | |
| Sweeden | 60-69 | **Gender** |
| United States | 70-79 | Male |
| United Kingdom | 80-89 | Female |

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 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: [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 5 versus 1 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: [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 vs introduced threat, a personal vs impersonal act, and harming one as a means vs a byproduct. 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; J. Mikhail, 2000) in order to reduce the number of relevant parameters or distinctions to potentially only one: means vs byproducts. 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 byproduct 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 = .3219). Because of the substantially smaller effect size, a sample of 420 subjects was necessary to achieve statistical power of .95. Employing this stringent criteria, we were able to test a small range of demographic subsets for the predicted dissociation in judgments: males, females, subjects aged 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 aged 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-byproduct 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. But 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 as 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.

## *2.3. 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 et al., 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. Like Denise, the man in the jeep has the option of watching the herd run over and kill five people, or drive 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 show that 12 out of 15 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 is not only 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 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.

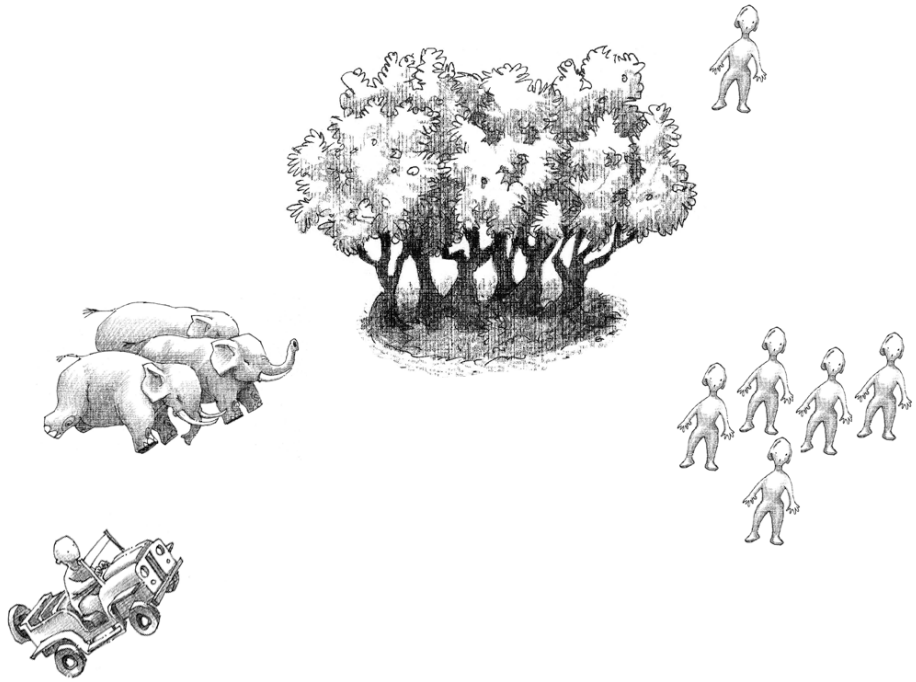

Figure 7. A content manipulation of the familiar Bystander trolley problem, designed for field testing among hunter-gather populations. Here, a man in a jeep has an opportunity to drive toward the herd of stampeding elephants, causing them to move around the grove, saving the five but killing the one person.

Patient populations present a particularly unique opportunity, especially when damage to an area or circuit results in selective deficits as opposed to more general cognitive dysfunction. Consider, for example, the case of autism. When Daniel Dennett (1978; 1983) originally raised the problem of mental state attribution — or having a theory of mind in David Premack's (1978) terminology — he simultaneously linked questions of beliefs, desires and intentions with the moral psychology of action. He pulled off this integration by asking the reader to imagine a classic Punch and Judy interaction. Having once again abused Punch, Judy walks away and then accidentally trips and falls into a box. Punch, seeing the opportunity for ending his abusive relationship, plans to push Judy off the stage, ending not only her career but her life. As he prepares his murderous act, he turns around to get some rope. With his back turned, Judy opens the box and sneaks off. Punch returns, ties the box up, shoves it off stage, and rejoices over his perfect homicide. Depending on your ability to mind read or simulate what is going on in Punch and Judy's heads, there are two interpretations. Punch believed that Judy was trapped inside the box. When he pushes the box off stage, he believes that he has killed her. Punch's belief is, however, false. If children only pay attention to consequences — as Piaget and Kohlberg believed was characteristic of children up to about nine years old — then Punch didn't do anything wrong. Judy is fine. If children pay attention to beliefs and intentions, then Punch did do something wrong. His plan failed, but this is not what he intended and it is not what he believes. If we had children of different ages on a jury, how would they vote? It turns out that children up to the age of about 4 fail to understand that Punch has a false belief. They believe that since they saw Judy escape, so

too did Punch. Punch therefore didn't do anything wrong. He was striking at thin air, an action in jest. In contrast, older children understand that they know something that Punch does not, and therefore, that he is smacking the box because he believes he will kill Judy. Even though there are no negative consequences, his intentions were clear. Autistic children typically live a life that is frozen at the under-4 years stage, never understanding that someone else can have different and false beliefs based on what they have or have not seen. Moreover, these individuals often fail to distinguish intentional from accidental acts and thus, tend to focus on consequences alone.

Given this characterization of the autistic/Asperger disorder, we predicted that those affected by it would respond differently from normal subjects on the MST in just those cases where the causal-intentional aspects of the case really make a difference when it comes to judging an agent's actions. To test this hypothesis, we linked our web site to those designed to provide information to autistics. To date, we have collected data from approximately 80 subjects clinically diagnosed with autism or Asperger's [high functioning autistics], focusing primarily on the family of trolley problems. If the intentional aspects matter, as we believe they do, then subjects should see little or no difference between any of the cases, focusing exclusively on consequences. Since action in each of these cases results in the same consequences — saving five, but killing one — each should be perceived as permissible. Results confirmed this hypothesis. First, our sample of autistics responded identically to normals with respect to the standard bystander scenario, viewing Denise's flipping the switch as permissible. Second, and in striking contrast to normals, 40% of autistics judged Frank's pushing the man as permissible. Third, and in contrast to normals, our sample of autistics perceived no difference between Ned and Oscar, two scenarios in which the only difference lies in terms of the mental states mediating the consequences: Ned uses the heavy man as a means to saving five, whereas Oscar kills the one as an incidental byproduct of his intent to stop the trolley by using the weight. Lastly, and in contrast with normals, a significantly greater proportion of autistics judged Ned's action as permissible; we note, however, that with the limited test population of autistic subjects our statistical power to detect the predicted dissociation between Ned and Oscar was only .41. Overall then, in the absence of access to the causal structure of actions and their consequences, many autistics perceive these moral dilemmas in a different light. Consequences appear to dominate because autistics are functionally blind to the parametric variation in actions mediated by different intentional states.

Returning to our four toy models, the autistic data highlight the significance of the appraisal system, and the importance of a Rawlsian creature. For autistics, part of the appraisal system is damaged. In the absence of an ability to distinguish intended as opposed to merely foreseen consequences, autistics focus strictly on consequences. Although these differences are striking, they also serve to raise two other points. First, even autistics perceive some difference between these dilemmas. For example, although they are more likely than normals to judge Frank's action as permissible, Frank's action is nonetheless perceived as less permissible than any of the other dilemmas. This residual may well be due to the emotional or personal aspect entailed in Frank's pushing the man. Consequently, some aspect of the emotional content of this dilemma is getting through and causing at least some perceptible differences between these cases. This kind of result, if supported by additional tests, would provide evidence that both an appraisal system operating over the causes and consequences of action, together with an emotional coloring, play a role in guiding moral judgments. In essence, both Rawlsian and Humean creatures would have a voice in generating a moral judgment. Second, though this component of the appraisal system plays a critical role in

generating moral judgments, it is not specific to the moral faculty. As discussed in the first section, any dissection of our moral faculty must assess whether the mechanism in play is specific to morality and uniquely human. The distinction between intended and foreseen consequences plays out in many non-moral situations [e.g., putting on music because you intend to have a party for which people come over, and doing the same for your personal pleasure, and having people nonetheless come over for a party having heard the music] and it is certainly possible that this kind of distinction is recognized by other animals given recent work on the intentional-accidental distinction (Call, Hare, Carpenter, & Tomasello, 2004; Hauser, Chen, Chen, & Chuang, 2003). In sum, recognizing the means to particular ends is, as Kant and many others recognized, critical to our moral faculty but it is not specialized for it.

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 part one, *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 non-patients 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 and Daniel Tranel, 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. Like autistics, these frontal lobe patients were more likely than normals to judge Frank as permissible, and to see no difference between Ned and Oscar, again suggesting that they are attending quite specifically to consequences. In the case of these patients, however, it is the apparent lack of emotional input that drives the focus on consequences. In this sense, damage to these frontal areas creates the ultimate utilitarian. Let us flesh this out by highlighting some of the other dilemmas.

We presented a series of four dilemmas, all derived from the logic of two classic Peter Singer cases. In each of these cases, an agent is presented with an opportunity to help, but the pull varies depending upon the scenario.

> Scenario 5: John, a white American, has just bought a brand new convertible sports car with white leather interior. He decides to take it for a spin on a country road. As he is driving along, he sees up ahead a man lying by the side of the road. John can see that one of the man's legs is covered in blood and he is clearly in pain. The man, of African descent, looks up at John and asks in a foreign accent, "Please sir, could you drive me to the hospital a few minutes down the road?" If John stops and brings the man to the hospital, the doctors will be able to stop the bleeding, thereby saving his leg. But if

John takes the man to the hospital, the leather interior of his car will be ruined, which will cost him $200 to repair.

Is it morally permissible for John to drive on, leaving the man behind?

Scenario 6: Timothy, a white American, has just walked into his house and finds a stack of mail. He sits down and begins putting the mail into three categories: trash, bills to pay, and personal. One letter is from UNICEF, an organization dedicated to the survival, protection, and development of children world wide. Timothy opens the letter and finds a card with three emaciated children on the cover. Inside, the card reads: "Thousands of children die each year in Saharan Africa due to lack of water. A contribution of $50 will save 25 lives by providing each child with a package of oral rehydration salts that will eliminate dehydrating diarrhea, and allow them to survive. We kindly ask that you contribute to saving 25 children from dying."

Is it morally permissible for Timothy to throw away the UNICEF card without contributing?

Scenario 7: Wes is walking through a crowded park on a cold winter evening. He is nearly home when he sees a homeless man. The man has no winter clothing, and soon he will freeze and die. Wes is wearing a warm coat that he could give to the man, saving his life. If Wes keeps his coat, the homeless man will freeze and die. If Wes gives the homeless man his coat, the homeless man will survive.

Is it morally permissible for Wes to keep his coat?

Scenario 8: Neil is walking through a crowded park on a cold winter evening. He is nearly home when he sees collection station for donations to the homeless. A sign explains that most homeless people have no winter clothing, and that dozens will freeze and die every night in the winter. Neil is wearing a warm coat that he could put in the collection station, saving the life of one homeless person. If Neil keeps his coat, a homeless person will freeze and die. If Neil puts his coat in the collection station, a homeless person will survive.

Is it morally permissible for Neil to keep his coat?

Most people say that it is not permissible for John to drive on, but it is permissible for Timothy to throw away the UNICEF card. Although frontal lobe patients see a difference between these cases, with more subjects judging Timothy's act as more permissible than John's, they tend to see both cases as less permissible than normals. This same pattern carries over to Wes and Neil. For both groups, it is more permissible for Neil to keep his coat than Wes, but the frontal patients see both actions as impermissible. Together, this class of dilemmas suggest that in the absence of a certain kind of emotional input, consequentialism rises as the default perspective. Importantly, though these patients look different from normals on these cases, for other dilemmas they don't. Ongoing work is aimed at fleshing out the nature of this disorder.

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 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.

## *3.0. Sweet justice! Rawls and 21st 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 1960-1980. 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 et al., 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.

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' linguistic analogy is clearly targeted at the descriptive level, even though many of his critics considered him to be saying more (J. Mikhail, 2000; Mikhail, in prep). 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 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' 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. Evidence that this is the case comes 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 practioners 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 ultimately and rationally works out to be the better policy from the standpoint of human welfare. But 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' 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' deep insight about the nature of our moral psychology, an instance of sweet justice[11].

## *4.0. References*


Anderson, S. R., & Lightfoot, D. (2000). The human language faculty as an organ. *Annual Review of Physiology, 62*, 697-722.

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

Baron, J. (1998). *Judgment Misguided: Intuition and error in public decision making.* Oxford: Oxford University press.

Baron, J., & Siepmann, M. (2000). Using web questionnaires for judgment and decision making research. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet.* (pp. 235-265). New York: Academic Press.

Bloom, P. (2004). *Descartes' Baby*. New York: Basic Books.

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). Do chimpanzees discriminate between an individual who is unwilling to share and one who is unable to share? *Developmental Science, xxx*, yyy-zzz.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Hillsdale: Lawrence Erlbaum.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* New York, NY: Praeger.

Chomsky, N. (1988). *Language and Problems of Knowledge.* Cambridge: MIT Press.

Chomsky, N. (2000). *On Nature and Language*. New York: Cambridge Univesity Press.

Damasio, A. (1994). *Descartes' Error.* Boston, MA: Norton.

Damasio, A. (2000). *The Feeling of What Happens.* New York: Basic Books.

Dehaene, S. (1997). *The Number Sense.* Oxford: Oxford University Press.

Dennett, D. C. (1978). Belliefs about beliefs. *Behavioral and Brain Sciences, 1*, 568-570.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: the 'Panglossian paradigm' defended. *Behavioral and Brain Sciences, 6*, 343-390.

Dwyer, S. (2004). *How good is the linguistic analogy*. Retrieved February 25, 2004, from www.umbc.edu/philosophy/dwyer

Fischer, J. M., & Ravizza, M. (1992). *Ethics: Problems and Principles*. New York: Holt, Rinehart & Winston.

Fitch, W. T., Hauser, M. D., & Chomsky, N. (in press). The evolution of the language faculty: clarifications and implications. *Cognition, xxx*, yyy-zzz.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5-15.

Frohlich, N., & Oppenheimer, J. A. (1993). *Choosing Justice: An Experimental Approach to Ethical Theory.* Berkeley: University of California Press.

Gert, B. (1998). *Morality: Its Nature and Justification.* New York: Oxford University Press.

Gert, B. (2004). *Common Morality.* New York: Oxford University Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Greenwald, A. G., Nosek, B., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: 1. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197-216.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814-834.
Harman, G. (1977). *The Nature of Morality: An Introduction to Ethics.* New York: Oxford University Press.
Hauser, M. D. (in press). *Moral Minds: the unconscious voice of right and wrong.* New York: Harper Collins.
Hauser, M. D., Chen, M. K., Chen, F., & Chuang, E. (2003). Give unto others: genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back. *Proceedings of the Royal Society, London, B, 270*, 2363-2370.
Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569-1579.
Hauser, M. D., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (in review). A dissociation between moral judgment and justification. *Proceedings of the National Academy of Sciences, USA, xxx*, yyy-zzz.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies.* New York: Oxford University Press.
Jackendoff, R. (2002). *Foundations of Language*. New York: Oxford University Press.
Jackendoff, R. (2005). *Language, Culture, Consciousness: Essays on Mental Structure.* Cambridge: MIT Press.
Kagan, S. (1988). The additive fallacy. *Ethics, 90*, 5-31.
Kamm, F. M. (1998a). Moral intuitions, cognitive psychology, and the harming-versus-not-aiding distinction. *Ethics, 108*, 463-488.
Kamm, F. M. (1998b). *Morality, Mortality: Death and whom to save from it.* New York: Oxford University Press.
Kraut, R., Olson, J., Banaji, M. R., Bruckman, A., Cohen, J., & Cooper, M. (2004). Psychological research online. *American Psychologist, February/March*, 105-117.
Lerdahl, F., & Jackendoff, R. (1996). *A Generative Theory of Tonal Music.* Cambridge: MIT Press.
Macnamara, J. (1990). The development of moral reasoning and the foundations of geomery. *Journal for the Theory of Social Behavior, 21*, 125-150.
Mikhail, J. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice'.* Unpublished PhD, Cornell University, Ithaca.
Mikhail, J. (in prep). *Rawls' linguistic analogy.* New York: Cambridge University Press.
Mikhail, J., Sorrentino, C., & Spelke, E. S. (1998). *Toward a universal moral grammar.* Paper presented at the Proceedings of the Cognitive Science Society.
Mikhail, J. M. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A theory of justice'.* Unpublished PhD, Cornell University, Ithaca.
O'Neill, P., & Petrinovich, L. (1998). A preliminary cross cultural study of moral intuitions. *Evolution and Human Behavior, 19*, 349-367.
Petrinovich, L., O'Neill, P., & Jorgensen, M. J. (1993). An empirical study of moral intuitions: towards an evolutionary ethics. *Ethology and Sociobiology, 64*, 467-478.
Pinker, S. (1994). *The Language Instinct.* New York: William Morrow and Company, Inc.

Pizarro, D., & Bloom, P. (2003). The intelligence of the moral emotions: a comment on Haidt (2001). *Psychological Review, 110*, 293-296.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 4*, 515-526.
Rachels, J. (1975). Active and passive euthanasia. *New Enlgand Journal of Medicine, 292*, 78-80.
Rawls, J. (1971). *A Theory of Justice.* Cambridge: Harvard University Press.
Saporta, S. (1978). An interview with Noam Chomsky. *Linguistic Analysis, 4[4]*, 301-319.
Schmidt, W. C. (1997). World-Wide-Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods and Computers, 29*, 274-279.
Smith, A. (1759/1976). *The Theory of the Moral Sentiments.* Oxford: Clarendon Press.
Spelke, E. S. (1994). Initial knowledge: six suggestions. *Cognition, 50*, 431-445.
Thomson, J. J. (1970). Individuating actions. *Journal of Philosophy, 68*, 774-781.
Tranel, D., Bechara, A., & Damasio, A. (2000). Decision making and the somatic marker hypothesis. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences* (pp. 1047-1061). Cambridge: MIT Press.
Wilson, E. O. (1975). *Sociobiology: A New Synthesis.* Cambridge: Harvard University Press.
Wilson, E. O. (1998). The biological basis of morality. *The Atlantic Monthly, April*, 53-70.

---

## *5.0. Footnotes*

[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. But 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, 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 [p. 72-73] "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." Further on, he states [p. 107] 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. " See Mikhail for a more comprehensive discussion of Rawls' linguistic analogy, together with several important extensions (J. M. Mikhail, 2000).

[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 the chapter, when we use the terms right, wrong, permissible and so forth, we are using these as shorthand for *morally* right, wrong, 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], etc; 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 that justify their answers by saying that killing is wrong, may have a more nuanced view concerning cause and effect, seeing the 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 9 patients with frontal damage. The extent and location of damage is quite variable between patients, including differences in whether it is unilateral or bilateral. The aim is to test a much larger sample and look more carefully at patterns of responses relative to patterns of damage with the additional help of Dan Tranel and Hanna Damasio.

[10] Though not made explicitly in this chapter, 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] 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' 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 chapter; 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.