This paper appeared in *The Origin of Values*, ed. by Michael Hechter, Lynn Nadel & Richard E. Michod. New York: Aldine de Gruyter. 1993. Pp. 215-228.

Moral Philosophy and Mental Representation

Stephen Stich
Department of Philosophy and Center for Cognitive Science
Rutgers University
New Brunswick, NJ 08901
stich@ruccs.rutgers.edu

Contents

Let me begin with a bit of autobiography. I am, by profession, a teacher of philosophy. Year in and year out, for the last fifteen or twenty years, I have taught a large undergraduate course on contemporary moral issues - issues like abortion, euthanasia, reverse discrimination, genetic engineering, and animal rights. Over the years, I have written a handful of papers on some of these topics. However, most of my research and writing has been in a very different domain. It has been concerned with problems in the philosophy of language, the philosophy of mind and the philosophy of psychology. During the last decade, much of my work has been on the philosophical foundations of cognitive science, and I have spent a great deal of time thinking and writing about the nature of mental representation.

For a long time I assumed that the two branches of my professional life were quite distinct. However, a few years ago I began to suspect that there might actually be important connections between them. The invitation to participate in the Tucson conference on the Scientific Analysis of Values has provided the motivation to set out my suspicions a bit more systematically. In reading what follows, do keep in mind that it is very much a first stab at these matters. I suspect that much of what I have to say is seriously oversimplified, and no doubt some of it is muddled or mistaken.

Here's an overview of what is to come. In Sections I and II, I will sketch two of the projects frequently pursued by moral philosophers, and the methods typically invoked in those projects. I'll argue that these projects presuppose (or at least suggest) a particular sort of account of the mental representation of human value systems, since the methods make sense only if we assume a certain kind of story about how the human mind stores information about values. The burden of my argument in Section III will be that, while the jury is still out, there is some evidence suggesting that this account of mental representation is mistaken. If it is mistaken, it follows that two of the central methods of moral philosophy may have to be substantially modified, or perhaps abandoned, and that the goals philosophers have sought to achieve with

these methods may themselves be misguided. I fear that many of my philosophical colleagues will find this a quite radical suggestion. But if anything is clear in this area, it is that the methods we will be considering have not been conspicuously successful, though it certainly hasn't been for want of trying. So perhaps it is time for some radical, empirically informed rethinking of goals and methods in these parts of moral philosophy.

In Section IV, I'll take a brief look at a rather different project in moral philosophy. This project, I'll argue, is compatible with a wide range of theories about the structures subserving mental representation. But to pursue the project seriously, it will be necessary determine which of these theories is correct. And that is a job requiring input from anthropologists, linguists, AI researchers and cognitive psychologists as well as philosophers. If this is right, a surprising redrawing of traditional disciplinary boundaries is in order. For a central project in ethics will turn out to be located squarely within the domain of cognitive science.

I. Plato's Quest: The Analysis of Moral Concepts

> Well said, Cephalus, I replied; but as concerning justice, what is it? - to speak the truth and to pay your debts - no more than this? And even to this are there not exceptions?

With this passage in the *Republic*,[(1)]

Plato launches a long inquiry whose goal is to find the definition of justice. Let me pick up the quote where I left off, since the next few sentences provide a paradigm for the process of inquiry Plato will pursue.

> Suppose that a friend when in his right mind has deposited arms with me and he asks for them when he is not in his right mind, ought I to give them back to him? No one would say that I ought or that I should be right in doing so, any more than they would say that I ought always to speak the truth to one who is in his condition.
>
> You are quite right, he replied.
>
> But then, I said, speaking the truth and paying your debts is not a correct definition of justice.
>
> Quite correct, Socrates....[(2)]

Much the same pattern recurs frequently in Plato's dialogues. Here's another example.

> *Socrates*. I abjure you to tell me the nature of piety and impiety, which you say that you know so well, and of murder, and of other offenses against the gods. What are they? Is not piety in every action always the same?
>
> *Euthyphro*. To be sure, Socrates.
>
> *Socrates*. And what is piety, and what is impiety? .... Tell me what is the nature of this idea, and then I shall have a standard to which I may look, and by which I may measure actions, whether yours or those of any one else, and then I shall be able to say that such and such an action is pious, such another impious.
>
> *Euthyphro*. I will tell you, if you like.... Piety ... is that which is dear to the gods, and impiety is that which is not dear to them.

> *Socrates*. Very good, Euthyphro; you have now given me the sort of answer which I wanted. But whether what you say is true or not I cannot as yet tell....
>
> The quarrels of the gods, noble Euthyphro, when they occur, are of a like nature [to the quarrels of men].... They have differences of opinion ... about good and evil, just and unjust, honorable and dishonorable....
>
> *Euthyphro*. You are quite right.
>
> *Socrates*. Then, my friend, I remark with surprise that you have not answered the question which I asked. For I certainly did not ask you to tell me what action is both pious and impious: but now it would seem that what is loved by the gods is also hated by them.[(3)]

Throughout the history of philosophy, there has been no shortage of authors who have followed in Plato's footsteps, seeking definitions of such central moral notions as justice, goodness, obligation, responsibility, equality, fairness and a host of others. Typically, those pursuing these projects share with Plato a cluster of assumptions about how the game is to be played. The first is that a correct definition must provide *individually necessary and jointly sufficient conditions* for the application of the concept being defined. It must specify what every instance falling under the concept, and only these, have in common. If there are exceptions to the definition - either cases which fit the definition but to which the concept does not apply, or cases which don't fit the definition to which the concept does apply - then the definition is mistaken.

A second widely shared Platonic assumption is that we already have a great deal of knowledge relevant to the definition we seek. The central strategy in testing a proposed definition is to compare what the definition says to what *we* would say about a variety of actual and hypothetical cases. On the definition offered by Cephalus, justice requires paying your debts. But we would not say that a man is unjust if he refuses to return the weapons of a friend who is no longer in his right mind. So Cephalus's definition must be mistaken. To make this sort of test work we must suppose that we already know whether or not refusing to return the arms would be unjust - we must have this sort of knowledge prior to articulating the sought after definition. Indeed, the Platonic inquiry seems to make the most sense if we assume that we already know necessary and sufficient conditions for the application of the concept, and that this knowledge is being put to work in guiding our judgements about the various cases, both real and hypothetical, that are offered as potential counter-examples to proposed definitions. Though of course at the beginning of the inquiry our knowledge of necessary and sufficient conditions is largely tacit; it is not available in a form that enables us to specify those conditions. If it were, the Platonic quest for definitions would be much easier than it is.

A third assumption underlying the Platonic project is that it will do some good to articulate and make explicit the necessary and sufficient conditions that, presumably, we already tacitly know. Socrates motivates his request for a definition by saying that when he has it "then I shall have a standard to which I may look, and by which I may measure actions,... and then I shall be able to say that such and such an action is pious, such another impious." There is something of a paradox lurking here, however. For, as we've just seen, the method that Plato and the many who follow him invoke seems to require that we already know how to "measure actions ... and say that such and such an action" is just or pious or what have you. Judgements about the applicability of the term we are seeking to define are the *input* into the process of testing proposed definitions. Having noted this paradox, I don't propose to pursue it any further, since doing so would take us too far afield.

II. Morally Relevant Difference Arguments

My second example of a project in moral philosophy that seems to make some strong assumptions about the mental representation of values is one that I find myself pursuing over and over again in my courses in contemporary moral issues. To motivate the project for my students, I begin with the observation that if two cases are to be judged differently from a moral point of view - if, for example, one action is judged morally right while another is morally wrong - then it must be the case that there is some non-moral feature with respect to which they differ. Two cases which are *exactly* the same in every descriptive or non-moral respect must be morally the same as well. Philosophers like to make this point by saying that the moral properties of a situation *supervene* on the non-moral properties. Once the former have been determined, the latter are fixed as well.

Now, by itself, this principle of the supervenience of the moral on the non-moral can't do much work for us, since in the real world there are no two cases that are exactly alike. There will always be *some* differences between any two situations. However, if we are going to draw a moral distinction between a pair of cases, the descriptive differences between them must be differences that we take to be *morally relevant* - they must be aspects of the situation that we are seriously prepared to accept as justifying the drawing of a moral distinction. And if they justify the drawing of a moral distinction in the case at hand, then presumably they justify the drawing of a parallel moral distinction in other cases that differ in the same way.

All of this will be a bit clearer if we consider an example. The illustration I'll use is one of my favorites in the classroom - the issue of animal rights. I begin the discussion by noting that most people have reasonably stable and reasonably clear views about what is right and wrong in this domain. The goal I propose to the students is the apparently modest one of making their own views explicit.

Most students are not vegetarians. They are prepared to say that there is nothing morally wrong with the practice of raising and slaughtering a variety of agricultural animals, for no better reason than that some people like to eat the meat of those animals. There is, in particular, nothing at all morally wrong with raising pigs destined for slaughter and ultimately for pork chops and ham sandwiches. Non-vegetarian students typically do not condone the *cruel* treatment of farm animals. And, of course, some of the most powerful arguments of animal rights advocates turn on what are alleged to be the intrinsically cruel nature of modern farming methods. But for the purposes of the current illustration, let us leave the issue of cruelty to one side. Let's assume that the animals we're considering are treated well and slaughtered as painlessly as possible. Under these circumstances most of my students are prepared to agree, indeed insist, that there is nothing wrong with raising cows, pigs and other common farm animals for food.

Now consider a parallel case. Suppose a group of very wealthy gourmets decide that it would pleasant to dine occasionally on human flesh. To achieve their goal they hire a number of couples who are prepared to bear infants to be harvested for the table. Typically, my students' first reaction to this proposal is horror and disgust, accompanied with considerable moral indignation. They are quite certain that such a practice would be morally intolerable. Very well, then, I ask, what is the morally relevant difference between farming children and farming animals? Why do you draw a moral distinction between babies and pigs? To start the ball rolling, I note that there are all sorts of features that distinguish adult humans from pigs which they can't appeal to here. It is not the case that a human baby is more intelligent than an adult pig, or more self-conscious, or more rational, or more aware of its environment. With respect to all of these features, adult pigs are *superior* to babies.

Well, the answer usually comes back, perhaps it's true that an adult pig is more intelligent and self aware than a human infant. But the difference is one of *potential*. Human babies have the potential to become significantly more intelligent, rational, self-aware, etc. than any pig can ever be. Human babies grow up to be moral agents. Pigs don't. And it is the potential for developing in these ways that marks the moral boundary between pigs and babies.

Ah, I reply, not so fast. Let me change the case a bit. Suppose that our gourmets, sensitive to concerns about potential, have arranged to treat the sperm with which the women are impregnated. The treatment makes some small changes in the genetic make-up of the sperm, with the result that the children produced are all very severely retarded. None of these children have the potential for developing into rational, reflective adults. On any reasonable measure, none of them will ever be as rational as a normal adult pig. Or, if you prefer, we can imagine yet another variation on the theme. Suppose our gourmets have entered into an arrangement with the administration of several large hospitals. Whenever there is an extremely senile patient in one of the hospitals who has no close relatives or friends, the patient is turned over to the gourmets, and ends up in the stew at their next banquet. Here again, there is no potential for rationality, or for becoming a moral agent. The people who end up on the dinner table have less potential, along these lines, than a normal adult pig.

At this point the students are generally getting a bit uncomfortable, and it is common for someone to propose that the crucial difference between the pig and the senile person or the severely retarded child is simply that the latter two are *humans* - they are members of *our* species. It is the difference between humans and non-humans that marks a major moral boundary. A first response to this suggestion, one that often comes from another student, is the observation that this is *speciesism* - a doctrine that bears a distressing similarity to racism. But if a student is unmoved by the analogy, the following tale will typically be very unsettling. Suppose it were to be found that some small group of people living amongst us - people of Icelandic descent, for example - turn out not to be able to have children when married to partners outside their group. On further investigation it turns out that the Icelanders are incapable of interbreeding with the rest of us because they are actually genetically different from us. They have a different number of chromosomes, and a significantly different genetic structure. They are, in short, members of a different species. The difference went unnoticed for so long because Icelanders generally marry other Icelanders. Despite the differences, however, Icelanders typically make exemplary citizens, and they are often the best of friends with non-Icelanders. Some of them do superb science, others write first rate poetry, and a fair number of them are skilled at sports. Nonetheless, they are members of another species. And because of the difference in species, our gourmets conclude they are morally justified in having the occasional Icelander for dinner - as the main course.

Not at all surprisingly, the students find this morally repugnant, and they concede that mere difference in species is not enough to mark the moral boundary they seek. Indeed, what often happens at this point in the discussion is that students start to question their initial moral judgments. If it is so *hard* to specify the morally relevant difference between pigs and babies, perhaps that is because there are no differences that they are prepared to take seriously. Perhaps when the issue at hand is killing for food, pigs and babies should not be treated differently. Perhaps what we do to pigs is horribly *wrong*. It is not at all uncommon for students to suffer a small moral crisis when confronted with these considerations. And in at least a few cases students who came back to visit a number of years later have told me that they had been strict vegetarians ever since taking my course.

The search for morally relevant differences between harvesting pigs and harvesting people is in some ways quite different from the Platonic search for definitions. In Plato's project, we are seeking to characterize conditions for the application of a particular moral notion like justice or responsibility or piety. In debating the morality of using animals for food, we are seeking the characterize an important moral boundary - the boundary between those creatures that it is morally acceptable to kill simply to satisfy our own tastes, and those which it would be morally repugnant to kill for this reason. However, there are also some important similarities between the two endeavors. In both investigations, we are trying to specify the extension of categories by seeking necessary and sufficient conditions. We want an account that will cleanly divide cases into two distinct classes - the just and the unjust, or the things it is permissible to kill for food and the things it is not permissible to kill for food. Also, it seems that in both cases we must assume that we already know a great deal about the categories we are seeking to

characterize. The methods proceed by testing proposed conditions against our "intuitive" judgments about actual and hypothetical cases. And, as we noted earlier, it is plausible to suppose that if this process is to succeed we must already have something like a set of tacitly known necessary and sufficient conditions to guide our judgments about particular cases. Thus, both the Platonic quest for definitions and the search for morally relevant differences appear to presuppose a view about the process underlying our ability to classify items into categories: *Categorization exploits tacitly known necessary and sufficient conditions*. In the section that follows, I will sketch some of the reasons to suspect that this account of categorization may be mistaken.

III. Categorization and Concepts

In the psychological literature, the cognitive structures underlying categorical judgements are generally referred to as *concepts*.[(4)] And in psychology, as in philosophy, there is a long-standing tradition which insists that concepts must specify necessary and sufficient conditions. However, since the early 1970s there has been a growing body of experimental literature challenging this "classical view" of concepts. Perhaps the most well known work in this area has been done by Eleanor Rosch and her associates.

In one series of experiments it was shown that people can reliably order instances falling under a concept when asked how "typical" or "representative" the instances are. Thus, for example, an apple will be rated as a more typical fruit than a lemon; a lemon will be rated as more typical than a coconut; and a coconut will be rated as more typical than an olive (Mervis, Catlin & Rosch, 1976; Rosch, 1978, Malt & Smith, 1984). What is important about these ratings is that they predict performance on a wide variety of tasks including categorization.

If subjects are asked whether a particular item is or is not a fruit, and are told to respond as quickly as possible, their responses are faster for more typical instances and slower for less typical instances (Smith, Shobin & Rips, 1974). Also, when subjects are asked to generate examples of sub-categories of a given concept they mention typical ones before atypical ones. Thus, for example, subjects asked to name kinds of fruit will mention apples, peaches and pears before blueberries, and blueberries will be mentioned before avocados or pumpkins (Rosch, 1978).

Now if a concept is the cognitive structure that subjects are using when they make categorical judgments, then these results begin to make the classical view of concepts look a bit problematic. For if concepts specify necessary and sufficient conditions, they apply equally to every instance of the concept, and it is not obvious why some instances should be more typical, easier to categorize and easier to recall.

Another line of research that has been taken to undermine the classical view of concepts suggests that typicality effects can be explained by appeal to properties that are common in members of the category, though they are not necessary conditions for membership in the category. In a number of studies, subjects were provided with a list of instances or sub-categories falling under a given concept, and they were asked to specify properties of the items on the list. Thus, for example, if the category in question is *birds*, subjects will be given a list that includes *robin, canary, vulture, chicken* and *penguin*. The properties that subjects offer for *canary* might include *has feathers, flies, small size, sings,* etc. Only the first two of these would be offered for *vulture*, and only the first for *penguin*. Given these data, we can compute what Rosch and her associates call the *family resemblance score* for various kinds of birds. This is done by assigning to each property (*has feathers, flies*, etc.) a number proportional to the number of bird kinds for which the property was mentioned. (Thus the number assigned to *has feathers* would be higher than the number assigned to *sings*. Having weighted the properties, the family resemblance score for a particular kind of bird is simply the sum of the weights of the properties mentioned for that bird. The high family resemblance score for *robin* indicates that robins have many properties that occur frequently in other

kinds of birds, while the low family resemblance score for *chicken* and *penguin* indicates that these birds do not have many of the properties that occur frequently in other sorts of birds. Not surprisingly, the family resemblance score turns out to be an excellent predictor of typicality, and thus an excellent predictor of categorization speed, recall, etc.

In the light of these results, a number of investigators have proposed accounts of concepts that are at odds with the classical (necessary and sufficient conditions) view. One widely discussed idea is that a concept consists of a set of salient features or properties which characterize only the best or "prototypical" members of category. This prototype representation will, of course, contain a variety of properties that are lacking in some members of the category. The prototypical bird flies, but emus don't. On the prototype view of concepts, objects are classified as members of a category if they are sufficiently similar to the prototype - that is, if they have a sufficient number of properties specified in the prototype representation. The more similar an item is to the target prototype, the faster one can determine that it exceeds the similarity threshold. Thus a more typical member of a category will be recognized and classified more rapidly than a less typical one.

Another proposal for dealing with the experimental results posits the mental representation of one or more specific exemplars. An exemplar is a specific instance of an item falling under a concept - the spaniel that was my boyhood companion (for *dog*) or the couch in our living room (for *couch*). On this view, categorization proceeds by activating the mental representations of one or more exemplars for the concept at hand, and then assessing the similarity between the exemplars and the item to be categorized. When developed in detail, exemplar models and prototype models yield different predictions, and there are some sophisticated empirical studies aimed at determining which model is superior in various conceptual domains (see, for example, Estes, 1986).

Recent research strongly suggests that neither the prototype approach nor the exemplar approach can tell the whole story about conceptual representation, even for simple object concepts like *fruit* and *bird* (Medin & Smith, 1984; Smith, 1990). The consensus seems to be that conceptual representation is a complex affair combining prototypes or exemplars with less observationally salient, more theoretical features. Also, it may well turn out that conceptual representation works differently in different domains. If this is right, then the mental representation of "goal derived" categories, like *things not to eat on a diet* and social concepts like *extrovert* or *communist* may have a format that is quite different from the mental representation of *apple*, *fruit*, or *dog* (Barsalou, 1987).

While the empirical story about the mental structures underlying categorization is far from complete, it should be clear that much of the work I have been reviewing poses a major challenge to the two methods in moral philosophy sketched in Sections I and II. For both of those methods assume that categorization exploits tacitly known necessary and sufficient conditions, and much of the empirical work on categorization suggests that classical necessary and sufficient conditions play little or no role in the process. To the best of my knowledge, there have been no empirical studies aimed at exploring the mental representation of moral concepts like *justice* or *responsibility*. Nor has anyone looked carefully at the cognitive structures underlying our ability to use categories like *things it is morally acceptable to kill for food*. However, if the story for those concepts is at all like the story elsewhere, it will explain why it is that moral philosophers working with the methods I sketched have been so unsuccessful for so long. For if the mental representation of moral concepts is similar to the mental representation of other concepts that have been studied, then the tacitly known necessary and sufficient conditions that moral philosophers are seeking *do not exist*.

Exemplar models of conceptual representation, and more sophisticated variations on the theme which invoke "scripts" or stories, also suggest an explanation for the fact that those engaged in moral pedagogy generally prefer examples to explicit principles or definitions. Myths, parables, fables, snippets of

biography (real or fanciful) - these seem to be the principle tools of a successful moral teacher. Perhaps this is because moral knowledge is *stored* in the form of examples and stories. It may well be that moral doctrines cast in the form of necessary and sufficient conditions are didactically ineffective because they are presented in a form that the mind cannot readily use.

IV. Some Alternative Models for the Mental Representation of Moral Systems

The two projects in moral philosophy that we've looked at so far seem to presuppose that the mental structures underlying moral judgments are rather like definitions - they specify individually necessary and jointly sufficient conditions for the application of moral concepts. The psychological models that challenge this presupposition offer alternative accounts of conceptual representation, accounts that don't involve necessary and sufficient conditions. But these alternatives are still very much *like* definitions. Indeed, as Quine pointed out long ago, a typical dictionary definition of a word like *tiger* or *lemon* will not offer necessary and sufficient conditions. Often it will present a list of features of a typical tiger or a typical lemon - very much in the spirit of the prototype account of mental representation (Quine, 1953).

However, in the philosophical literature there is a venerable tradition that suggests a rather different account of how moral judgments are made. Instead of relying on something akin to definitions, this tradition assumes that our moral judgments are derived from an interconnected set of *rules* or *principles* specifying what sort of actions are just or unjust, permissible or not permissible, and so on. There are some clever ways in which certain systems of rules can be recast as a set of necessary and sufficient conditions. Thus the distinction between these two approaches is not a hard and fast one. Still, in many cases the style and complexity of rule-based theories give them a very different appearance and a very different feel.

In his seminal book, *A Theory of Justice*, John Rawls (1971, 46) urges that a first goal of moral philosophy should be the discovery of the set of rules or principles underlying our reflective moral judgment. (46) These principles along with our beliefs about the circumstances of specific cases, should entail the intuitive judgments we would be inclined to make about the cases, at least in those instances where our judgments are clear, and there are no extraneous factors likely to be influencing them. There is, of course, no reason to suppose that the principles guiding our moral judgments are fully (or even partially) available to conscious introspection. To uncover them we must collect a wide range of intuitions about specific cases (real or hypothetical) and attempt to construct a system of principles that will entail them.

As Rawls notes, this method for uncovering the system of principles presumed to underlie our intuitive moral judgments is analogous to the method used in modern linguistics. Following Chomsky, linguists typically assume that speakers of a natural language have internalized a system of generative grammatical rules, and that these rules play a central role in language production and comprehension. The rules are also assumed to play a central role in the production of linguistic intuitions - the more or less spontaneous judgments speakers offer about the grammaticality and other linguistic properties of sentences presented to them. In attempting to discover what a speaker has internalized, linguists construct systems of generative rules, and check them against the speaker's intuitions. However, the internalized rules are not the only psychological system that plays a part in producing reported intuitions. Memory, motivation, attention and other factors all interact in the production of the judgments speakers offer. Thus the rules the linguist produces should not be expected to capture the exact details of the speakers' judgments. As in the case of moral principles, we expect the rules to capture only the clearest intuitions, and even these may be ignored when there is some reason to suspect that other factors are distorting the subject's judgment.

Now, as Rawls (1971: 47) observes, it is very likely that the grammatical rules for a natural language like English will "require theoretical constructions that far outrun the ad hoc precepts of our explicit grammatical knowledge." So if the analogy between grammar and ethics is a good one, "there is no reason to assume that our sense of justice can be adequately characterized by familiar common sense precepts" (Rawls, 1971: 47). It may also be the case that the principles underlying our moral intuitions, like those underlying our grammatical intuitions, are both numerous and enormously complex. Indeed, in the case of language, Chomsky has long maintained that the rules are so complex that they could not possibly be learned from the relatively limited data available to the child. Rather, he contends, the range of grammars that it is possible for a child to learn is a small and highly structured subset of the set of logically possible grammars. Thus much of the fundamental structure of the grammars that children ultimately internalize must be innate. One of the more intriguing possibilities suggested by the analogy between grammatical theory and moral theory is that, as we learn more about the mental representations underlying moral judgement, we may find that they sustain a similar sort of "argument from the poverty of the stimulus". Thus it may be that "humanly possible" moral systems are a very small subset of the logically possible systems, and that much of the structure of moral systems is innate, not acquired.

Though grammatical knowledge was one of the first domains to be systematically investigated by cognitive scientists, there has been a great deal of work on the mental structures underlying other sorts of knowledge, belief and skill during the last two decades. Mathematical knowledge, knowledge of various sciences, and common sense knowledge in various domains have all been explored. The cognitive systems underlying various skills, from chess and computer programming to musical composition and medical diagnosis, have also been investigated. Theories attempting to account for people's abilities in these areas have invoked a wide range of knowledge representing systems, some of them rather like the generative systems that loom large in linguistics, and others quite different. What makes this work relevant to our current concerns is that many of the knowledge or belief systems that have been explored are at least roughly analogous to moral systems. In many cases people can offer a complex, subtle and apparently systematic array of judgments about particular cases, with little or no conscious access to the mechanisms or principles underlying these judgments. Thus, while Rawls was certainly right in the noting parallels between ethics and grammar, there are other analogies that are at least as plausible. Perhaps the mental structures underlying moral judgment are similar to those underlying expert medical diagnosis, or commonsense physical intuition. Perhaps the best analogy is with the knowledge structures that guide our expectations in reading stories about restaurants and other common social situations.

Which account of the mental representation of moral systems is best is certainly not a matter to be settled a priori. The question is an empirical one. But it is, I think, the sort of empirical question that is best approached in a resolutely interdisciplinary way. Philosophers have lavished a great deal of attention on the exploration of moral intuitions, and have amassed a very substantial body of cases illustrating the richness, subtlety and complexity of our moral judgments. Linguists and deontic logicians have studied the semantic and logical structure of moral language. Anthropologists have much to say about the moral systems in cultures very different from our own. AI researchers, particularly those concerned with knowledge representation, have explored the strengths and weaknesses of many strategies for storing and using complex bodies of information. And, of course, cognitive psychologists have a sophisticated bag of tricks for testing hypotheses about the form and content of mentally represented information.

It is my strong suspicion that progress in understanding how people represent and use moral systems will not be made until scientists and scholars from these various disciplines begin to address the problem collaboratively. Indeed, one of my goals in writing this paper is to convince at least some of my readers that it is high time to launch such a collaborative effort.

A final note: If I am right about the way to make headway in understanding how moral systems are mentally represented, and if Rawls is right in suggesting that such an understanding is a first essential step

in moral philosophy, then the beginnings of moral philosophy fall squarely within the domain of cognitive science.

References:

Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser, ed., *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press.

Estes, W. (1986). Array models for category learning. *Cognitive Psychology*, 18.

Malt, B. & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, 23.

Medin, D. & Smith, E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35.

Mervis, C., Catlin, J., & Rosch, E. (1976). Category structure and the development of categorization. In R. Spiro, B. Bruce and W. Brewer, eds., *Theoretical Issues in Reading Comprehension*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Plato (1892). *The Dialogues of Plato*, translated by B. Jowett. Vol. I. New York: Random House.

Quine, W. (1953). Two dogmas of empiricism. In *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rey, G. (1983). Concepts and stereotypes. *Cognition*, 15.

Rey, G. (1985). Concepts and conceptions: A reply to Smith, Medin & Rips. *Cognition*, 19.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd, eds., *Cognition and Categorization*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Smith, E. (1989). Three distinctions about concepts and categorization. *Language and Mind*, 4, 1 & 2.

Smith, E. (1990). Categorization. In D. Osherson & E. Smith, eds., *Thinking: An Invitation to Cognitive Science*, Vol. 3. Cambridge, MA: MIT Press.

Smith, E., Medin, D., & Rips, L. (1984). A psychological approach to concepts: Comments on Rey's 'Concepts and stereotypes.' *Cognition*, 17.

Smith, E., Shoben, E., and Rips, L. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81.

Notes

1. Plato (1892), The *Republic* I, 331, p. 595.

2. Plato (1892), The *Republic* I, 331, p. 595.

3. Plato (1892*), Euthyphro*, 5-7, pp. 386-389.

4. Philosophers too sometimes use the term 'concept' in this way, though they also use the term in some very different ways. For a useful discussion of the contrast, see Rey (1983) & (1985), Smith, Medin & Rips (1984), Smith (1989).