

Mind and Morals: Essays on Ethics and  
Cognitive Science

May, Friedman + Clark (eds)  
Cambridge, MA; MIT PRESS, 1996  
pp 109-127.

MAY, FRIEDMAN + CLARK (eds)  
MIND & MORALS

## Chapter 6

### Connectionism, Moral Cognition, and Collaborative Problem Solving

*Andy Clark*

How should linguistically formulated moral principles figure in an account of our moral understanding and practice? Do such principles lie at the very heart of moral reason (Kohlberg 1981)? Or do they constitute only a shallow, distortive gloss on a richer prototype-based moral understanding (Churchland 1989; Dreyfus and Dreyfus 1990)? The latter view has recently gained currency as part of a wider reassessment of the proper cognitive scientific image of human cognition—a reassessment rooted in the successes of a class of computational approaches known as connectionist, parallel distributed processing, or neural network models (McClelland, Rumelhart, and the PDP Research Group 1986). In this chapter I will argue that such approaches call not for the marginalization of the role of linguistically formulated moral rules and principles but for a thorough reconception of that role. This reconception reveals such summary maxims as the guides and signposts that enable collaborative moral exploration rather than as failed attempts to capture the rich structure of our individual moral knowledge. The force of this reconception eludes us, however, if we cast public linguistic exchange as primarily a tool for manipulating the moral understanding of other agents (Churchland 1989). Instead, we must focus on the role of such exchanges in attempts to engage in genuinely collaborative moral problem solving. A satisfying connectionist model of moral cognition will need to address the additional inner mechanisms by which such collaborative activity becomes possible and to recognize the ways in which such activity transforms the space of our moral possibilities.

#### *Connectionism: From Rules to Prototypes*

There is a conception of moral reason that informed many classical philosophical treatments but that has recently been called into question. At the heart of this conception lies a vision of informed moral choice as involving the isolation and application of an appropriate law or rule. Thus, to give a

simplicistic example, we might, on encountering some complex social situation, see it as falling under a rule prohibiting lying and act accordingly. Of course, the body of rules we are supposed (on this conception) to have internalized need not be so simple. Such a moral code could, as Ruth Barcan Marcus points out, be highly elaborate.<sup>1</sup> Principles could be hedged by exception clauses or rank ordered to help deal with potential conflicts. The logic of the moral code could be a fuzzy logic, or a nonmonotonic one, or something else. But however elaborate, the basic vision remains the same. It is a vision in which moral judgments are taken to involve "the judging of particular cases as falling under a particular moral concept, and thereby being governed by a specific moral rule" (Johnson 1993, 207). Johnson calls such a doctrine "moral law folk theory" (4–6). This doctrine, he claims, permeates our cultural heritage and hence underpins both lay and philosophical conceptions of the moral life. Yet it is a doctrine that, he argues, is radically mistaken, and indeed morally incorrect. It is, Johnson claims, "morally irresponsible to think and act as though we possess a universal, disembodied reason that generates absolute rules, decision-making procedures, and universal or categorical laws" (5).

Moral law folk theory is false, Johnson suggests, because it "pre-supposes a false account of the nature of human concepts and reason." This false account depicts concepts (including those that figure in putative moral rules and universals) as possessing classical structure. A concept possesses classical structure if it can be unpacked so as to reveal a set of necessary and sufficient conditions for its application. These necessary and sufficient conditions would effectively define the concept. To grasp the concept, on such a model, is to know that definition and to rely on it when called upon to deploy the concept. The trouble, as is now well known (Smith and Medin 1981; Rosch 1973) is that most (perhaps all) human concepts do not possess classical structure. Thus, whereas the classical model predicts that instances should fall squarely within or outside the scope of a given concept (according to whether the necessary and sufficient conditions are or are not met), robust experimental results reveal strong so-called typicality effects. Instances are classified as more or less falling under a concept or category according, it seems, to the perceived distance of the instance from prototypical cases. Thus, a dog is considered a better example of a pet than a tortoise, and a robin a better example of a bird than a pigeon. Such findings sit uneasily beside the classical image.<sup>2</sup> It seems we do not simply test for the presence or absence of a neat list of defining features and judge the concept applicable or inapplicable accordingly. In place of definitions and application rules invoking them, we face a vision of cognition organized around prototypical instances. To bring this vision into focus, we need to say a little more about the very idea of a prototype.

"Prototype" is sometimes used to mean merely a stereotypic example of the membership of some category. Thus understood, the stereotypic pet might be a dog, or the stereotypic crime a robbery. The recent popularity of prototype-invoking accounts in psychology and artificial intelligence, however, depends on a related but importantly different conception. Here, the notion of prototype is not the notion of a real, concrete exemplar. Rather, it is the notion of the statistical central tendency of a body of concrete exemplars. Such central tendency is calculated by treating each concrete example as a set of co-occurring features and generating (as the prototype) a kind of artificial exemplar that combines the statistically most salient features. Thus, the prototypical pet may include both dog and cat features, and the prototypical crime may include both harm to the person and loss of property. Concrete exemplars and rich worldly experience are still crucial, but they act as sources of data from which these artificial prototypes are constructed. Novel cases are then judged to fall under a concept (such as "pet" or "crime") according to the distance separating the set of features they exhibit from the prototypical feature complex—hence, the typicality effects mentioned earlier. (For a full review, see Clark 1993.)

Such a vision of prototype-based reason fits very satisfyingly with a particular model of information storage in the brain. This is the model of state-space representation, which draws on both neuroscientific conjecture and recent work with computational simulations of a broadly connectionist type (P. S. Churchland and T. J. Sejnowski 1992; P. M. Churchland 1989; McClelland, Rumelhart, and the PDP Research Group 1986, vols. 1, 2; Clark 1989, 1993). The main philosophical proponent of the state-space conception is undoubtedly Paul Churchland, who has also urged its importance for conceptions of moral reason (P. M. Churchland 1989, chap. 14). The flavor of the state-space approach is best conveyed by tracing out a simple example.

Consider the brain's representation of color. This representation (the example comes from P. M. Churchland 1989, 104) is fruitfully conceived as involving a three-dimensional (3D) state space (Land's color cube; see Land 1977) in which the dimensions (axes) reflect (1) long-wave reflectance, (2) medium-wave reflectance, and (3) short-wave reflectance. Each such dimension, Churchland conjectures, may correspond to the activity of downstream neural groups tuned to the activity of three different kinds of retinal cone. Within such a 3D space, white and black occupy diametrically opposed locations, while red and orange are quite close together. Our judgments concerning the perceived similarity-difference relations between colors may thus be explained as reflecting distance in this color-state space. The space thus exhibits what has been termed (Clark 1989, 1993) an inbuilt semantic metric.

Connectionist networks constitute one way of both implementing and acquiring representational spaces of this sort. Such networks consist of a complex of units (simple processing elements) and connections. The connections may be positive or negative valued (excitatory or inhibitory). The features of a stimulus are coded as numerical values (corresponding to the intensity of the feature or the degree to which it is present) across a designated group of units. These values are then differentially propagated through the network courtesy of the positive or negative connection weights. A good assignment of weights is one that ensures that activity in designated output units corresponds to some target input-output function.<sup>3</sup> Several layers of units may intervene between input and output. The activity of the units in each such layer will generally correspond to some kind of recoding of the input data that simplifies further processing. It is often fruitful to take each unit of such an intervening ("hidden") layer as one dimension of an acquired representational state space and to investigate the way the system responds to specific inputs by creating patterns of activity that define locations in this space (hidden unit activation space).

The great achievement of connectionism is to have discovered a set of learning rules that enables such systems to find their own assignments of weights to connections. The operation of such learning rules (which I shall not attempt to describe here, but see Clark 1989 and Churchland 1989 for accessible treatments) results in the construction of high-dimensional state spaces with associated semantic metrics. Four important features of this constructive process are:

1. It is exemplar-driven.
2. It is not bound by the similarity metric on the input vector.
3. It yields prototype-style representations.
4. It treats inference and reasoning as vectorial transformations across state spaces.

The learning process is exemplar driven in that the tuning of the weights is achieved by exposure to concrete cases. Thus, a network whose target is to transform written text into coding for phonemes (hence speech) does not have its weights changed by exposure to rules of text-phoneme conversion. Instead, it must be exposed to textual inputs, allowed to output an attempted coding for phonemes, and then amend its weights according to the difference between the target output and its actual performance.

The weight assignments that a net thus acquires can exploit hidden layers so as to dilate and compress the input space. Thus, two exemplars whose coding at the input layer is very similar may be pulled apart by the weights leading up the hidden layer. This is useful if, for example, two visual input descriptions are quite similar yet require very different

responses. Thus, two situations that are visually very similar (such as a person giving money to a beggar versus the same person giving money to a mugger holding a knife) may require very different responses. In such cases, the net can learn to use the hidden layer to recode the inputs in a new way such that the pattern of hidden unit activation is radically dissimilar in the cases just described. Correlatively, it may learn to code superficially dissimilar cases (such as giving money to a beggar and posting a check to a charity) with very similar hidden unit patterns. The state space defined by the hidden units might thus come to reflect a moral metric, whereas the input space depicted a visual one.

Within such state spaces, the mode of representation will come to exhibit features of prototype-style encodings. The reason is that features common to several training examples will figure in more episodes of weight adjustment than the less common features. As a result, the system will become especially adept at encoding and responding to such features. In addition, the learning regime will ensure that features that commonly occur together in the exemplars become strongly mutually associated. The upshot is that the system extracts the so-called central tendency of the body of exemplars, that is, a complex of common, co-occurring features. Moreover, multiple such complexes can be extracted and stored by a single net. McClelland and Rumelhart (1976) describe a net that (1) learns to recognize individual dogs by associating visual information with names, (2) extracts the central tendency of the body of dog exemplars, and hence exhibits knowledge of a prototypical set of dog features, and (3) can perform this trick for several different categories, simultaneously encoding knowledge about dogs, cats, and bagels in a single network. These various prototypes (of dog, cat, and bagel) are each coded for by distinct patterns of unit activation and hence determine different locations in a general state space whose dimensionality corresponds to the number of processing units. Individual dogs are coded by points relatively close to the dog prototype. Dogs and cats share more features with each other than do either with bagels; hence, the dog and cat prototypes lie relatively close together and at some fair remove from the bagel prototype. The system can use its knowledge of prototypical feature complexes to behave sensibly given novel inputs. To the extent that some new input exhibits several familiar features (for example, a half-dog/half-cat) the system will assign it to an appropriate location (in this case, midway between the dog and cat prototypes) and hence yield suitable outputs.

Reasoning and inference can now be reconstructed as processes of pattern completion and pattern extension. A network exposed to an input depicting the visual features of a red-spotted face may learn to activate a pattern of hidden units corresponding to a diagnosis of measles and a prescription of penicillin. The vector-to-vector transformation involved is

of a piece with that by which we perform simple acts of recognition and categorization such as naming a familiar dog. On the face of it, it is a million miles away from the intellectualist artificial intelligence model, which would have us consult a body of rules and principles and issue a medical judgment accordingly. (For a variety of similar examples and claims, see P. M. Churchland 1989, chap. 10.)

With this rough understanding of vector transformation models in place, we can now begin to address the issues concerning moral knowledge and reason. The primary lessons of the new approach, when applied to the moral domain, look to be twofold. First, the successful acquisition of moral knowledge may be heavily dependent on exposure not to abstractly formulated rules and principles but to concrete examples of moral judgment and behavior. (Literature, by depicting complex moral situations, may be seen as another kind of concrete case—virtual moral reality, if you will.) Second, our individual moral knowledge and reasoning may not be fully reconstructible in the linguistic space afforded by public language moral dialogue and discussion. This will be the case if, as seems likely, the internal representational space (or spaces) involved has even a fairly modest number of dimensions. Our sense of smell, as P. M. Churchland (1989, 106) notes, looks to involve at least a 6D space. If each dimension can take just ten different values, a space of  $10^6$  distinct locations is immediately available. Dog olfactory space, Churchland calculates, is of the order of  $30^7$  (22 billion) possible locations (compare to a world population of just 3.5 billion). State-space encoding thus allows even limited internal resources of units and weights to support representational spaces of great magnitude. Given the size of the brain's resources, the expressive capacity of biologically realistic inner systems looks unimaginably huge. The attempt to condense the moral expertise encoded by such a system into a set of rules and principles that can be economically expressed by a few sentences of public language may thus be wildly optimistic, akin to trying to reduce a dog's olfactory skills to a small body of prose.

These two implications (concerning the role of exemplars and the resistance of moral knowledge to summary linguistic expression) are remarked on by several recent writers.<sup>4</sup> Goldman (1993) notes the central role of exemplars, Churchland (1989) stresses in addition the general resistance of high-dimensional state-space encoded knowledge to linguistic expression, and Johnson (1993) describes how prototype-style encodings take us "beyond rules." The rule-based moral vision, according to this emerging consensus, is a doomed attempt to reconstruct the high-dimensional space of moral reason in a fundamentally low-dimensional medium. Such a diagnosis casts valuable light on questions concerning the rationality of moral thought. A well-tuned network, in command of state spaces of great complexity, may issue judgments that are by no means irrational but yet

resist quasi-deductive linguistic reconstruction as the conclusion of some moral argument that takes summary expressions of moral rules and principles as its premises. Such a vision is by no means new. Nagel comments that "the fact that one cannot say why a certain decision is the correct one ... does not mean that the claim to correctness is meaningless.... What makes this possible is *judgement* [which can] in many cases be relied on to take up the slack that remains beyond the limits of explicit rational argument" (Nagel 1987, 180), or again, "We know what is right in a particular case by what we may call an immediate judgement, or an intuitive subsumption ... moral judgements are not discursive" (Bradley 1876). As Bradley points out, phrases like "judgment" and "intuitive subsumption" are "perhaps not very luminous" (65). The value of the cognitive scientific excursion into state-space representation is just to cast a little light. It helps make concrete sense of a form of rational moral choice that nonetheless outruns what Nagel called "explicit rational argument."

The realization that individual moral know-how may resist expression in the form of any set of summary moral rules and principles is important. But it has mistakenly (or so I shall argue) led some writers to marginalize the role of such summary linguistic expressions in our moral life. It is this correlative marginalization that I now set out to resist. Such marginalization, I shall suggest, is the result of a common error: the error of seeing talk in general (and moral talk in particular) as primarily the attempt to reflect the fine-grained contents of individual thought. Such a view seems implicit in, for example, Paul Churchland's general skepticism concerning linguistic renditions. Such skepticism is evidenced in passages such as the following: "Any declarative sentence to which a speaker would give confident assent is merely a one-dimensional projection—through the compound lens of Wernicke's and Broca's areas onto the idiosyncratic surface of the speaker's language—a one-dimensional projection of a [high] dimensional solid that is an element in his true kinematical state" (P. M. Churchland 1989, 18). The high-dimensional solid is, of course, the internalized prototype-style know-how contained in a trained-up neural network. It is this know-how whose linguistic echo is but the flickering shadow on the wall of Plato's cave (Churchland 1989, 18).

More radically still, Dreyfus and Dreyfus (1990) go so far as to demote such low-dimensional linguaform projections to the status of mere tools for the novice—ladders to be kicked away by the true moral expert. Once again, the radical claim has some plausible roots in the observation that (as far as we can tell) truly expert ability (at chess, car driving, philosophy, moral reasoning) is not subserved by a set of compact rules or principles encoded quasi-linguistically by the brain. Instead, it is subserved by the operation of a fast, unreflective, connectionist-style resource or resources whose operation yields "everyday, intuitive ethical expertise" (246).

According to Dreyfus and Dreyfus, the expert, under normal conditions, “does not *deliberate*. She does not reason. She does not even act deliberately. She simply spontaneously does what has normally worked and, naturally, it normally works” (243).

This kind of fluid expertise comes, if it comes at all, only at the end of an extended learning history whose early stages are indeed marked by episodes of linguistic instruction. Dreyfus and Dreyfus in fact distinguish four stages that they claim precede fluid expertise: novice, advanced beginner, competence, and proficiency. Linguistic instruction figures prominently (unsurprisingly) in the initial novice stage, while linguistic reflection figures to a degree in all the other nonexpert stages: “The instruction process begins with the instructor decomposing the task environment into context-free features which the beginner can recognize without benefit of experience” (240). These context-free features are then used as components of rough-and-ready rules. Thus the would-be chess player is taught the numerical values of pieces (in context-free terms) and told to exchange pieces whenever a profit would accrue. Similarly, the would-be moral agent is told that to say intentionally what is false is to lie and that lying is in general to be avoided. (See Flanagan 1991 for a richer and more realistic treatment of this example.)

Dreyfus and Dreyfus are surely right to stress the role of language in novice learning. But they go on, wrongly I believe, to marginalize the role of language in truly expert behavior. They write that “principles and theories serve only for early stages of learning,” and as a result, “the skill development model we are proposing ... demotes rational, post-conventional moral activity to the status of a regression to a pre-expert stage of moral development” (252–256).<sup>5</sup> We are thus invited to treat linguistic justification and linguistically couched reflection as mere beginners’ tools—rough instruments not to be found in the tool kit of the true moral expert. I do not think this is the case. Rather, linguistic reflection and exchange enables a tuning and orchestration of moral response that is vital to moral expertise. What is needed is not a rejection of the role of summary linguistic expression and linguaform exchange in advanced moral cognition. Rather, we must reconceive that role. Such a reconception will occupy us for the remainder of this treatment.

#### *Language as a Manipulative Tool*

As a first move toward such a reconception, consider a question recently raised by P. M. Churchland. How, on a connectionist/prototype-based view does moral knowledge, once achieved, get modified and altered? Churchland bids us distinguish between the kinds of slow, experientially driven adaptive change and learning that configure the weights of individ-

ual networks over time and the kinds of fast flash-of-insight style “learning” that seems to occur when we suddenly see that a question with which we have been wrestling is easily solved once we reconceive the domain in the light of some new idea. The question Churchland raises is: How (if at all) is slow, connectionist-style learning to cope with fast flash-of-insight style conceptual change? The answer he develops is that it is able to do so because of the operation of so-called context fixers—additional inputs that are given alongside the regular input and that may cause an input that (alone) could not activate an existing prototype to in fact do so. Churchland terms such a process “conceptual redeployment” because it often leads to the reinvocation of prototypes developed in one domain in another, superficially very different, one.

Imagine someone trying to solve a problem. To solve it, if the approach outlined in the previous section is correct, is to activate an appropriate explanatory prototype. Sometimes, however, our attempts to access a satisfying (explanatory) prototype fail. One diagnosis may be that we do not command any appropriate prototype, in which case there is no alternative to slow, experience-based learning. But an alternative possibility is that we do command just such a prototype but have so far not called it up. This is where a good piece of context fixing can help. The idea is that a bare input that previously led to the activation of no fully explanatory prototype may suddenly, in the context of additional information, give rise to the activation of a developed and satisfying prototype by being led to exploit resources originally developed for a different purpose. Huygens, we are told, commanded a powerful wave prototype developed for water and sound media. Once he was led (by luck, scholarship, or something else) to combine questions regarding optics with context-fixing inputs concerning light, the optical questions were able to activate the rich and explanatory wave prototypes originally devised for the water domain. The conceptual revolution thus achieved did not involve slow, weight-adjustment-style learning, but rather consisted in “the unusual deployment of old resources” (23). The context-fixing information thus biases the treatment of an input vector in ways that can radically alter the prototype-invoking response of on-board, trained-up networks.

Now this, as Churchland notes, invites a certain perspective on linguaform debate, for linguistic exchanges can be seen as a means of providing fast, highly focused, context-fixing information. Such information may, as we have seen, induce others to activate prototypes they already command in situations in which those very prototypes would otherwise remain dormant. According to this view, moral debate does not work by attempting to trace out nomological-deductive arguments predicated on neat linguaform axioms. But summary moral rules and linguistic exchanges may nonetheless serve as context-fixing descriptions that prompt others to

activate certain stored prototypes in preference to others (see, e.g., comments in Churchland 1989, 300). Applying our story to an example from Johnson (1993), a moral debate may consist in the exchange of context fixers, some of which push us toward activation of an "invasion of privacy" prototype while others prompt us to conceptualize the very same situation in terms of a "prevention of espionage" prototype.

Note that according to such a vision the linguaform expressions do not aim to embody the reasoning that underlies individual moral judgment. Instead, they figure in exchanges whose goal is simply to prompt another's rich prototype-based knowledge to settle on one existing prototype rather than another. Thus, talk of "unborn children" may bias prototype-activation one way, while talk of "unwanted pregnancy" may bias it another. Moral rules and principles, on this account, are nothing more than one possible kind of context-fixing input among many. Others could include well-chosen images or non-rule-invoking discourse. Thus understood, language simply provides one fast and flexible means of manipulating activity within already developed prototype spaces. It is a simple matter, however, to extend this treatment to encompass a special role for summary principles, etc. in individual moral reflection. To see how, consider a nonmoral example case.

Kirsh and Maglio (1992, 1994) have investigated the roles of reaction and reflection in expert performance of the computer game Tetris in which the player attempts to accumulate a high score by the compact placement of geometric objects (Tetrazoids, or just Zoids) that fall down from the top of the screen. As a Zoid descends, the player can manipulate its fall by rotating it, moving it to the right or left, or instantly relocating it at the resting point of its current trajectory. When a Zoid comes to rest, a new one appears at the top of the screen. The speed of fall increases with score, and (the saving grace) a full row (one in which each screen location is filled by a Zoid) disappears entirely. When the player falls behind in Zoid placement and the screen fills up so that new Zoids cannot enter it, the game ends. Advanced play thus depends crucially on fast decision making. Hence, Tetris provides a clear case of a domain in which connectionist, pattern-completion style reasoning is required for expert performance. If the Dreyfus and Dreyfus model is correct, moreover, such parallel, pattern-completion style reasoning should exhaustively explain expert skill. But interestingly, this does not seem to be so. Instead, expert play looks to depend on a delicate and nonobvious interaction between a fast, pattern-completing module and a set of explicit, higher-level concerns or normative policies. The results are preliminary, and it would be inappropriate to report them in detail. But the key observation is that true Tetris experts report that they rely not solely on a set of fast, adaptive responses produced by, as it were, a trained-up network but also on a set of high-level

concerns or policies that they use to monitor the outputs of the skilled network so as to "discover trends or deviations from . . . normative policy" (Kirsh and Maglio 1992, 10). Examples of such policies include, "don't cluster in the center, but try to keep the contour flat" and "avoid piece dependencies" (Kirsh and Maglio 1992, 8–9). On the face of it, these are just the kind of rough-and-ready maxims that we might (following Dreyfus and Dreyfus) associate with novice players only. Yet attention to these normative policies seems to mark especially the play of real experts. Still, we must wonder how such policies can help at the level of expert play given the time constraints on responses. There is just no time for reflection on such policies to override online output for a given falling Zoid.

Here Kirsh and Maglio (1992) make a suggestive conjecture. The role of the high-level policies, they suggest, is probably indirect. Instead of using the policy to override the output of a trained-up network, the effect is to alter the focus of attention for subsequent inputs. The idea is that the trained-up network ("reactive module" as they put it) will sometimes make moves that lead to danger situations—situations in which the higher-level policies are not being reflected. The remedy is not to override the reactive module but thereafter to manipulate the inputs it receives so as to present feature vectors that, when processed by the reactive module in the usual way, will yield outputs in line with policy. As they describe it, the normative policies are the business of a distinct planner system that interacts rather indirectly with the online reactive agency: "It is the job of the planner to formulate a specification of concerns. These concerns are translated into directives for changing the focus of attention. Changes in attention in turn affect the feature vector presented to the [reactive agency]" (10). Just how the shift of attention is accomplished is left uncomfortably vague. But they speculate that it could work by "biasing certain board regions" or by "increasing the precision of [certain] values being returned by visual routines" (10).

Despite this vagueness, the general idea is attractive. Effective outputs are always under the control of the trained-up reactive system. But high-level reflection makes a contribution by effectively reconfiguring the input vectors that the reactive agencies receive.

This idea may provide a hint of a solution to the problem of understanding the role of explicitly formulated general commitments (in the form of summary rules or moral maxims) in moral thought.<sup>6</sup> Such commitments—the upshot of individual moral reflection—may help us monitor the outputs of our online, morally reactive agencies. When such outputs depart from those demanded by such policies, we may be led to focus attention on such aspects of input vectors as might help us bring our outputs back into line. Suppose we explicitly commit ourselves to an ideal

of acting compassionately in all circumstances. We then see ourselves reacting with anger and frustration at the apparent ingratitude of a sick friend. By spotting the local divergence between our ideal and our current practice, we may be able to bias our own way of taking the person's behavior—in effect, canceling out our representation of those aspects of the behavior rooted in their feelings of pain and impotence. To do so is to allow the natural operation of our on-board reactive agencies to conform more nearly to our guiding policy of compassion. The summary linguistic formulation, on this account, is a rough marker that we use to help monitor the behavior of our trained-up networks.

The moral of the Tetris example, then, is that advanced pattern recognition is really a double skill. In addition to the basic, fluent pattern-recognition-based responses exemplified by a trained connectionist net, the human expert relies on a second skill. This is the ability to spot cases in which these fluent responses are not serving her well. Such recognition (a kind of second-order pattern recognition) is crucial since it can pave the way for remedial action. And it is especially crucial in the moral domain. Here, surely, it is morally incumbent on us not to be hostage to our own fluent daily responses, no matter how well "trained" we are. We must be able to spot situations (for example, dealing with sexual politics in a family setting or interacting with certain religious or political groups) in which these fluent responses are failing to serve us. The effect of formulating some explicit maxims and guidelines provides us with a comparative resource in a sense external to our own online behaviors. This resource is neither binding nor a full expression of our moral knowledge, but it can act as a signpost alerting us to possible problems. The advanced moral agent, like the advanced Tetris player, needs to use every means available to sustain successful performance.

The cases just rehearsed go some way toward correcting the antilinguistic bias discerned in the previous section. Summary linguistic formulations, it seems, are not just tools for the novice. They are tools for the expert too. But the story remains sadly incomplete, for the image of linguistic tools suggests a merely manipulative role. This manipulative role does not, I claim, do justice to the more primary role of linguistic exchange as a medium of genuinely collaborative problem solving. Yet it is under this collaborative aspect (or so I shall argue) that linguistic formulations make their key contribution to moral cognition. It is to this perspective that we now turn.

#### *Language as a Collaborative Medium*

Missing from the discussion so far is any proper appreciation of the special role of language and summary moral maxims within a cooperative moral

community. To see this, we can begin by considering the general phenomenon of so-called collaborative learning. The observation here is simply that a procedure of multiple, cooperative perspective taking often allows groups of agents to solve problems that would otherwise defeat them. For example, two children, neither of whom is alone able to come to an understanding of the Piagetian conservation task (understanding how the same quantity of liquid can be manifest in very different ways in differently shaped vessels, such as a long, thin glass and a short, fat one) can often cooperate to solve the problem. The reason is that they "are often focussing on different aspects of the problem—one saying that the water in the new beaker is higher and the other noting it is thinner, for example. . . . These competing perspectives come to light in the interaction, and in an effort to reach a consensus the children integrate the perspectives, co-constructing, a new perspective" (Tomasello, Kruger, and Ratner 1993, 501; see also Perret-Clermont and Brossard 1985).

It is the communal effort to achieve consensus that drives the children to find the solution. Key features of this effort include discussion, joint planning, critiquing of each other's ideas, and requests for clarification. Many of these features are transactive in the sense of Kruger (1992). This means that the thinking and perspective of individual members of a group are objects of group attention and discussion. Given the crucial role of such modes of discussion, it is perhaps unsurprising to learn that collaborative learning emerges at about the same developmental moment (age six or seven) as does so-called second-order mental state talk—talk about other people's perspectives on your own and others' mental states. Thus, younger children (age three or four) are capable of seeing others as having a perspective on the world (seeing others as what Tomasello, Kruger, and Ratner 1993 call mental agents). But it is only the older children who see others as "reflective agents"—agents whose perspective includes a perspective on the child's own thought and cognition (see Tomasello, Kruger, and Ratner 1993, 501). Collaborative learning Tomasello, Kruger, and Ratner argue, requires a participant to recognize others as having ideas about each other's thoughts and perspectives. It requires participants to "understand in an integrated fashion the mental perspectives of two or more reflective agents" (501).

Such a capacity is plausibly viewed as an essential component of advanced moral cognition. Indeed, many moral problems basically consist in the need to find some practical way of accommodating multiple perspectives, including perspectives on each others' views and interests. Consider a typical moral issue such as how to accommodate the multiple, and often competing, perspectives and needs of different religions and racial groups in a multicultural society. Attempts to find practical solutions to the kinds of problems thus raised depend crucially on the extent to which

representatives of each group are able to engage in what may be termed multiple nested perspective taking. Consider the case of a conflict within a multicultural educational system.<sup>7</sup> The parents of a Muslim girl requested that she be excused from events involving what (from their perspective) was an unacceptably close physical proximity to boys. The head teacher was inclined to let the child decide. But the likely effect of the child's decision (she did not want to be excluded) would be her total removal from the school. In such a case, the only hope for a practicable solution lies in each party's willingness to try to understand the perspective of the other. It is here, I claim, that the role of linguistic exchange is paramount. The attempts by each party to articulate the basic principles and moral maxims that inform their perspective provide the only real hope of a negotiated solution. Such principles and maxims have their home precisely there: in the attempt to lay out some rough guides and signposts that constrain the space to be explored in the search for a cooperative solution. Of course, such summary rules and principles are themselves negotiable, but they provide the essential starting point of informed moral debate. Their role is to bootstrap us into a kind of simulation of the others' perspectives, which is, as we saw, the essential fodder of genuine collaborative problem-solving activity. No amount of such bootstrapping, of course, can preclude the possibility of genuine conflict between incompatible principles. But it is the exchange of such summary information that helps set the scene for the cooperative attempt to negotiate a practical solution to the problem at hand. Such a solution need not (and generally will not) consist in agreement on any set of general moral rules and principles. Instead, it will be a behavioral option tailored to the specific conflict encountered (see Khin Zaw, unpublished, for just such a defense of "practical reason").

Thus viewed, the rules and maxims articulated along the way are not themselves the determinants of any solution, nor need we pretend that they reveal the rich structure and nuances of the moral visions of those who articulate them. What they do reveal is, at best, an expertise in constructing the kinds of guides and signposts needed to orchestrate a practical solution sensitive to multiple needs and perspectives. This is not, however, to give such formulations a marginal or novice-bound role, nor is it to depict them as solely tools aimed at manipulating all parties into the activation of a common prototype. Rather, it is a matter of negotiating some practical response that accommodates a variety of competing prototypes. (The difference here is perhaps akin to that marked by Habermas's distinction between strategic and communicative action. In strategic action, the goal is to persuade the other, by whatever means, to endorse your viewpoint. In communicative action the goal is to motivate the other

to pursue a dialogue by visibly committing oneself to a *negotiated* solution. (Habermas 1990, 58, 59, 134, 145).<sup>8</sup>

The successful use of language as a medium of moral cooperation thus requires, it seems, an additional and special kind of knowing how—one not previously recognized in connectionist theorizing.<sup>9</sup> It concerns knowing how to use language so as to convey to others what they need to know to facilitate mutual perspective taking and collaborative problem solving. The true moral expert is often highly proficient at enabling cooperative moral debate. Moral expertise, *pace* Dreyfus and Dreyfus, cannot (for moral reasons) afford to be mute. This additional know-how, like the other expert skills discussed in the first section, may well itself consist in our commanding a certain kind of well-developed prototype space, but it will be a space that is interestingly second-order in that the prototypes populating it will need to concern the informational needs of other beings: beings who themselves can be assumed to command both a rich space of basic prototypes concerning the physical, social, and moral world and a space of second-order prototypes concerning ways to use language to maximize cooperative potential.

It is perhaps worth remarking, to emphasize the psychological reality of the complex of second-order skills, that high-functioning autistic children (those with basic linguistic skills) show a marked selective deficit in almost all of the areas I have discussed. It is characteristic of such children to show all of the following: no use of self-regulatory speech or inner rehearsal to help them perform a task (compare the Tetris example); very limited grasp of how to use language to achieve communicative goals; complete failure to recognize others as having a perspective on the child's own mental states; and no evidence of collaborative learning, or any other collaborative activity (Frith 1989, 130–145). These children, Frith suggests, are not able to "share with the listener a wider context of interaction in which both are actively involved" or to "gauge the comprehension of listeners" (126). They will use terms that no one else can understand, such as calling seventeen to twenty-five year olds the "student nurses age group" (125), and they "tend not to check whether their speech is actually succeeding and communicating, nor to they show any curiosity as to why a dialogue has broken down" (Baron-Cohen 1993, 512). The linguistic skills of these high-functioning autistics thus leave out all the collaborative dimensions I have been at pains to stress. As a result, Baron-Cohen (1993) raises the possibility that such children are, in a deep sense, acultural: unable to participate in the shared understanding and cooperative action essential to any true cultural group. Oversimplified connectionist models of moral cognition, by marginalizing the collaborative dimensions of moral action, likewise threaten to isolate the moral agent from her proper home, the moral community.

To sum up, it is only in the context of thinking about genuinely collaborative moral activity that the true power and value of principle-invoking moral discourse becomes visible. Summary moral rules and maxims act as flexible and negotiable constraints on collaborative action. Such rules and principles by no means exhaustively reflect our moral knowledge, but they are the expertly constructed guides and signposts that make possible the cooperative exploration of moral space.

#### *Conclusions: Complementary Perspectives on Moral Reason*

The kind of exchange between cognitive science and ethics that underlies the present treatment is quite typical. Historically, the bias of computational cognitive science is toward the individual. Ethical theory, by contrast, has concerned itself from the outset with individuals considered as parts of larger social and political wholes. The attempt to formulate a joint image of moral cognition helps correct the historical biases of each tradition. The ethicist is asked to think about the individual mechanisms of moral reason. The cognitive scientist is reminded that moral reason involves crucial collaborative, interpersonal dimensions. Perhaps neither party strictly requires the other to remind it of the neglected dimensions. But in practice, it is often the joint confrontation of the issues that yields progress in the search for an integrated image. In thus striving for a mutually satisfactory vision, we are forced to discover a common vocabulary and to agree on some focal issues, and to the extent that we do so, we prepare the ground for future participants from still other disciplines.

Such long-term benefits aside, the immediate upshot of this discussion is clear: recent connectionist-inspired reflections on moral cognition are probably right in asserting both that moral thinking is fruitfully depicted as a case of prototype-based reasoning and that summary linguistic principles and maxims can therefore provide only an impoverished gloss on the full complexities of our moral understanding. But the associated tendency to marginalize the role of such principles and maxims (to depict them as mere tools for the moral novice; Dreyfus and Dreyfus 1991) is to be resisted. As we saw, such formulations provide powerful tools for the indirect manipulation of moral cognition both in ourselves and others, and, most important, essential signposts and constraints that guide collaborative problem-solving activity. Such collaborative activity is only possible, I argued, courtesy of a special kind of knowing how: a knowing how whose focus is on the informational needs that must be met if others are to participate with us in cooperative problem-solving activity. Such know-how (knowing how to use language to prime the collaborative problem-solving machinery) requires a certain conception of other agents—a conception that recognizes others as already enjoying a particular perspective on the

thoughts and viewpoints of their fellows. In the light of all this, we can now see much that is missing from the basic connectionist story. A satisfying story about moral cognition and moral expertise must attend to a variety of thus far neglected, communication-specific, higher-order prototype spaces. To do so will be to recognize that the production and exploitation of summary linguistic rules and principles is not the production and exploitation of mere imperfect mirrors of moral knowledge. Rather, it is part and parcel of the very mechanism of moral reason.

#### *Acknowledgments*

I extend special thanks to Margaret Walker, Larry May, Marilyn Friedman, Owen Flanagan, Teri Mendelsohn, Peggy DesAutels, the members of the Washington University Ethics Seminar, the Philosophy/Neuroscience/Psychology work-in-progress group, and the audience at the 1993 Mind and Morals Conference at Washington University in St. Louis.

#### *Notes*

1. See Ruth Barcan Marcus, "Moral dilemmas and consistency," in C. W. Gowans (ed.), *Moral Dilemmas* (New York: Oxford University Press, 1987), pp. 188–204. The comments concerning the potential elaboration of the moral code occur on pp. 190–191.
2. "Uneasily," because the typicality findings are not conclusive evidence against a classical view. See Armstrong, Gleitman, and Gleitman (1983) and Osherson and Smith (1981).
3. Not all networks have designated output units, but the basic device of state-space representation characterizes the knowledge acquired even by so-called pattern association models.
4. "Summary," because extended treatments (such as those of classic literature) may indeed convey detailed information about the structure of moral space. "Summary linguistic expression" refers instead to attempts to distill moral knowledge into short rules and principles.
5. "Postconventional" here refers to stage 6 of Kohlberg's hierarchy of moral development (Kohlberg 1981), a stage at which principles are used to generate decisions.
6. I thank Peggy DesAutels for drawing my attention to the importance of such general normative commitments.
7. This is an actual case, borrowed from Susan Khin Zaw, "Locke and Multiculturalism: Toleration, Relativism and Reason," unpublished manuscript.
8. Habermas often assimilates the idea of strategic action to the idea of the manipulation of others by force or sanctions. Obviously, the idea of manipulation by provision of context-fixing input is importantly different. The question when such provision constitutes genuine manipulation as opposed to collaborative investigation is a delicate and important one. I note in passing that Habermas' emphases also echo those of this treatment in other ways, such as the recognition of the importance of multiple perspective taking (Habermas 1990, 138–146) and the conception of norms as practical, flexible aids rather than rigid defenses (180).
9. This was pointed out to me by Margaret Walker, whose help and comments have improved this chapter in numerous ways.

## References

- Armstrong, S., Gleitman, L., and Gleitman, H. 1983. "On What Some Concepts Might Not Be." *Cognition* 13:263-308.
- Baron-Cohen, S. 1993. "Are Children with Autism Acultural?" *Behavioral and Brain Sciences* 16:512-513.
- Bradley, F. H. 1876. "Collision of Duties." In C. W. Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press, 1987.
- Churchland, P. M. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, Mass.: MIT Press.
- Churchland, P. M. Forthcoming. "Learning and Conceptual Change: The View from the Neurons." In A. Clark and P. Millican, eds., *Essays in Honour of Alan Turing*. Oxford: Oxford University Press.
- Churchland, P. S., and Sejnowski, T. J. 1992. *The Computational Brain*. Cambridge, Mass.: MIT Press.
- Clark, A. 1989. *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*. Cambridge, Mass.: MIT Press.
- Clark, A. 1993. *Associative Engines: Connectionism, Concepts and Representational Change*. Cambridge, Mass.: MIT Press.
- Dreyfus, H., and Dreyfus, S. 1990. "What is Morality? A Phenomenological Account of the Development of Ethical Expertise." In D. Rasmussen, ed., *Universalism vs. Communitarianism: Contemporary Debates in Ethics*. Cambridge, Mass.: MIT Press.
- Flanagan, O. 1991. *Varieties of Moral Personality: Ethics and Psychological Realism*. Cambridge, Mass.: Harvard University Press.
- Frith, U. 1989. *Autism*. Oxford: Blackwell.
- Goldman, A. 1993. "Ethics and Cognitive Science." *Ethics* 103:337-360.
- Habermas, J. 1990. *Moral Consciousness and Communicative Action*, translated by C. Lenhardt and S. Weber Nicholsen. Cambridge, Mass.: MIT Press.
- Johnson, M. 1993. *Moral Imagination: Implications of Cognitive Science for Ethics*. Chicago: University of Chicago Press.
- Khin Zaw, S. n.d. "Does Practical Philosophy Rest on a Mistake?" Unpublished manuscript.
- Kirsh, D., and Maglio, P. 1992. "Reaction and Reflection in Tetris." In J. Hendler, ed., *Artificial Intelligence Planning Systems: Proceedings of the First Annual International Conference AIPS 92*. San Mateo, Calif.: Morgan Kaufman.
- Kirsh, D., and Maglio, P. 1994. "On Distinguishing Epistemic from Pragmatic Action." *Cognitive Science* 18:513-549.
- Kohlberg, L. 1981. *Essays on Moral Development*. Vol. 1, *The Philosophy of Moral Development*. New York: Harper & Row.
- Kruger, A. C. 1992. "The Effect of Peer and Adult-Child Transaction Discussions on Moral Reasoning." *Merill-Palmer Quarterly* 38:191-211.
- Land, E. 1977. "The Retinex Theory of Color Vision." *Scientific American* (December): 108-128.
- McClelland, J., and Rumelhart, D. 1976. "A Distributed Model of Human Learning and Memory." In J. McClelland, D. Rumelhart, and the PDP Research Group, *Parallel Distributed Processing*. Cambridge, Mass.: MIT Press.
- McClelland, J., Rumelhart, D., and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. 2 vols. Cambridge, Mass.: MIT Press.
- Marcus, Barcan R. 1987. "Moral Dilemmas and Consistency." In C. W. Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press.
- Nagel, T. 1987. "The Fragmentation of Value." In C. W. Gowans, ed., *Moral Dilemmas*. New York: Oxford University Press.

- Osherson, D., and Smith, E. 1981. "On the Adequacy of Prototype Theory as a Theory of Concepts." *Cognition* 9:35-38.
- Perret-Clermont, A. N., and Brossard, A. 1985. "On the Interdigitation of Social and Cognitive Processes." In R. A. Hinde, A. N. Perret-Clermont, and J. Stevenson-Hinde, eds. *Social Relationship and Cognitive Development*. Clarendon Press, Oxford.
- Rosch, E. 1973. "Natural Categories." *Cognitive Psychology* 4:324-350.
- Sejnowski, T., and Rosenberg, C. 1987. "Parallel Networks That Learn to Pronounce English Text." *Complex Systems* 1:145-168.
- Smith, E., and Medin, D. 1981. *Categories and Concepts*. Cambridge, Mass.: Harvard University Press.
- Tomasello, M., Kruger, A., and Ratner, H. 1993. "Cultural Learning." *Behavioral and Brain Sciences* 16:495-552.