

edited by

Larry May, Marilyn Friedman,  
and Andy Clark

# mind and morals

virtue  
caution  
modesty  
decision  
judgment

Essays on Ethics and Cognitive Science

Second printing, 1998

© 1996 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Palatino by Asco Trade Typesetting Ltd., Hong Kong and was printed and bound by in the United States of America.

Library of Congress Cataloging-in-Publication Data

Mind and morals : essays on cognitive science and ethics / edited by  
Larry May, Marilyn Friedman, and Andy Clark.

p. cm.

"A Bradford book."

Includes bibliographical references and index.

ISBN 0-262-13313-X (alk. paper). — ISBN 0-262-63165-2

(pbk. : alk. paper)

I. Cognitive science—Moral and ethical aspects. 2. Cognitive science.

3. Ethics. I. May, Larry. II. Friedman, Marilyn, 1945—

III. Clark, Andy.

BJ45.5.M56 1995

170—dc20

95-22361

CIP

## Contents

Acknowledgments vii

Contributors ix

Chapter 1

Introduction 1

*Larry May, Marilyn Friedman, and Andy Clark*

Part I

Ethics Naturalized? 17

Chapter 2

Ethics Naturalized: Ethics as Human Ecology 19

*Owen Flanagan*

Chapter 3

How Moral Psychology Changes Moral Theory 45

*Mark L. Johnson*

Chapter 4

Whose Agenda? Ethics versus Cognitive Science 69

*Virginia Held*

Part II

Moral Judgments, Representations, and Prototypes 89

Chapter 5

The Neural Representation of the Social World 91

*Paul M. Churchland*

Chapter 6

Connectionism, Moral Cognition, and Collaborative Problem

Solving 109

*Andy Clark*

## Chapter 5

# The Neural Representation of the Social World

*Paul M. Churchland*

---

### *Social Space*

A crab lives in a submarine space of rocks and open sand and hidden recesses. A ground squirrel lives in a space of bolt holes and branching tunnels and leaf-lined bedrooms. A human occupies a physical space of comparable complexity, but in our case it is overwhelmingly obvious that we live also in an intricate space of obligations, duties, entitlements, prohibitions, appointments, debts, affections, insults, allies, contracts, enemies, infatuations, compromises, mutual love, legitimate expectations, and collective ideals. Learning the structure of this social space, learning to recognize the current position of oneself and others within it, and learning to navigate one's way through that space without personal or social destruction is at least as important to any human as learning the counterpart skills for purely physical space.

This is not to slight the squirrels and crabs, or the bees and ants and termites either, come to think of it. The social dimensions of their cognitive lives, if simpler than ours, are still intricate and no doubt of comparable importance to them. What is important, at all levels of the phylogenetic scale, is that each creature lives in a world not just of physical objects but of other creatures as well, creatures that can perceive and plan and act, both for and against one's interests. Those other creatures therefore bear systematic attention. Even nonsocial animals must learn to perceive, and to respond to, the threat of predators or the opportunity for prey. Social animals must learn, in addition, the interactive culture that structures their collective life. This means that their nervous systems must learn to represent the many dimensions of the local social space, a space that embeds them as surely and as relevantly as does the local physical space. They must learn a hierarchy of categories for social agents, events, positions, configurations, and processes. They must learn to recognize instances of those many categories through the veil of degraded inputs, chronic ambiguity, and the occasional deliberate deception. Above all, they must learn to generate appropriate behavioral outputs in that social

space, just as surely as they must learn to locomote, grasp food, and find shelter.

In confronting these additional necessities, a social creature must use the same sorts of neuronal resources and coding strategies that it uses for its representation of the sheerly physical world. The job may be special, but the tools available are the same. The creature must configure the many millions of synaptic connection strengths within its brain so as to represent the structure of the social reality in which it lives. Further, it must learn to generate sequences of neuronal activation-patterns that will produce socially acceptable or socially advantageous behavioral outputs. As we will see in what follows, social and moral reality is also the province of the physical brain. Social and moral cognition, social and moral behavior, are no less activities of the brain than is any other kind of cognition or behavior. We need to confront this fact, squarely and forthrightly, if we are ever to understand our own moral natures. We need to confront it if we are ever to deal both effectively and humanely with our too-frequent social pathologies. And we need to confront it if we are ever to realize our full social and moral potential.

Inevitably these sentiments will evoke discomfort in some readers, as if, by being located in the purely physical brain, social and moral knowledge were about to be devalued in some way. Let me say, most emphatically, that devaluation is not my purpose. As I see it, social and moral comprehension has just as much right to the term "knowledge" as does scientific or theoretical comprehension—no more right, but no less. In the case of gregarious creatures such as humans, social and moral understanding is as hard won, is as robustly empirical and objective, and is as vital to our well-being as is any piece of scientific knowledge. It also shows progress over time, both within an individual's lifetime and over the course of many centuries. It adjusts itself steadily to the pressures of cruel experience, and it is drawn ever forward by the hope of a surer peace, a more fruitful commerce, and a deeper enlightenment.

Beyond these brief remarks, the philosophical defense of moral realism must find another occasion. With the patient reader fairly forewarned, let us put this issue aside for now and approach the focal issue of how social and moral knowledge, whatever its metaphysical status, might actually be embodied in the brains of living biological creatures.

It cannot be too difficult. Ants and bees live intricate social lives, but their neural resources are minuscule—for an ant,  $10^4$  neurons, tops. However tiny those resources may be, evidently they are adequate. A worker ant's neural network learns to recognize a wide variety of socially relevant things: pheromonal trail markings to be pursued or avoided; a vocabulary of antennae exchanges to steer one another's behavior; the occasions for general defense, or attack, or fission of the colony; fertile pasture for the

nest's aphid herd; the complex needs of the queen and her developing eggs; and so forth.

Presumably the challenge of social cognition and social behavior is not fundamentally different from that of physical cognition and behavior. The social features or processes to be discriminated may be subtle and complex, but as recent research with artificial neural networks illustrates, a high-dimensional vectorial representation—that is, a complex pattern of activation levels across a large population of neurons—can successfully capture all of them. To see how this might be so, let us start with a simple case: the principal emotional states as they are displayed in human faces.

### EMPATH: A Network for Recognizing Human Emotions

Neural net researchers have recently succeeded in modeling some elementary examples of social perception. I here draw on the work of Garrison Cottrell and Janet Metcalfe at the University of California, San Diego. Their three-stage artificial network is schematically portrayed in figure 5.1.

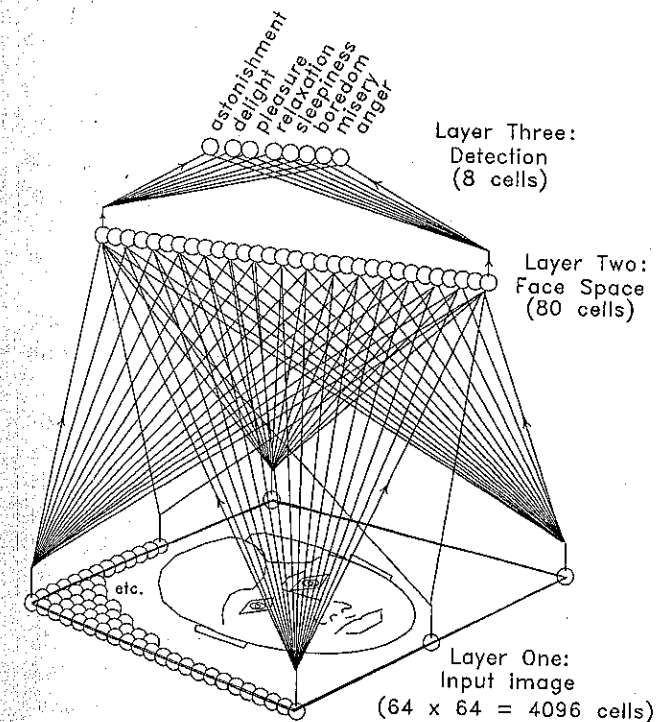


Figure 5.1  
EMPATH, a feedforward network for recognizing eight salient human emotions.



Figure 5.2  
Eight familiar emotional states, as feigned in the facial expressions of three human subjects. From the left, they are astonishment, delight, pleasure, relaxation, sleepiness, boredom, misery, and anger. These photos, and those for seventeen other human subjects, were used to train EMPATH, a network for discriminating emotions as they are displayed in human faces.

Its input layer or "retina" is a  $64 \times 64$ -pixel grid whose elements each admit of 256 different levels of activation or "brightness." This resolution, both in space and in brightness, is adequate to code recognizable representations of real faces.

Each input cell projects an axonal end-branch to every cell at the second layer of 80 cells. That layer represents an abstract space of 80 dimensions in which the input faces are explicitly coded. This second layer projects finally to an output layer of only 8 cells. These output cells have the job of explicitly representing the specific emotional expression present in the current input photograph. In all, the network contains  $(64 \times 64) + 80 + 8 = 4,184$  cells, and a grand total of 328,320 synaptic connections.

Cottrell and Metcalfe trained this network on eight familiar emotional states, as they were willingly feigned in the cooperating faces of twenty undergraduate subjects, ten male and ten female. Three of these charming subjects are displayed eight times in figure 5.2, one for each of the eight emotions. In sequence, you will there see astonishment, delight, pleasure, relaxation, sleepiness, boredom, misery, and anger. The aim was to discover if a network of the modest size at issue could learn to discriminate features at this level of subtlety, across a real diversity of human faces.

The answer is yes, but it must be qualified. On the training set of (8 emotions  $\times$  20 faces =) 160 photos in all, the network reached—after 1000 presentations of the entire training set, with incremental synaptic adjustments after each presentation—high levels of accuracy on the four positive emotions (about 80 percent), but very poor levels on the negative emotions, with the sole exception of anger, which was correctly identified 85 percent of the time.

Withal, it did learn. And it did generalize successfully to photographs of people it had never seen before. Its performance was robustly accurate for five of the eight emotions, and its weakest performance parallels a similar performance weakness in humans. (Subsequent testing on real humans showed that they too had trouble discriminating sleepiness, boredom, and misery, as displayed in the training photographs. Look again at figure 5.2, and you will appreciate the problem.) This means that the emotional expressions at issue are indeed within the grasp of a neural network, and it indicates that a larger network and a larger training set might do a great deal better. EMPATH is an "existence proof," if you like: a proof that for some networks, and for some socially relevant human behaviors, the one can learn to discriminate the other. Examination of the activation patterns produced at the second layer—by the presentation of any particular face—reveals that the network has developed a set of eight different prototypical activation patterns, one for each of the eight emotions it has learned, although the patterns for the three problematic negative emotions are diffuse and rather indistinct. These eight prototypical patterns are what the final eight output units are tuned to detect.

#### *Social Features and Prototypical Sequences*

EMPATH's level of sophistication is, of course, quite low. The "patterns" to which it has become tuned are timeless snapshots. It has no grasp of any expressive sequences. In stark contrast to a normal human, it will recognize sadness in a series of heaving sobs no more reliably than in a single photograph of one slice of that telltale sequence. For both the human and the network, a single photograph might be ambiguous, but to the human, that distressing sequence of behavior certainly will not be. Lacking any recurrent pathways, EMPATH cannot tap into the rich palette of information contained in how perceivable patterns unfold in time. For this reason, no network with a purely feed-forward architecture, no matter how large, could ever equal the recognitional capacities of a human.

Lacking any grasp of temporal patterns carries a further price. EMPATH has no conception of what sorts of causal antecedents typically produce the principal emotions, and no conception of what effects those emotions have on the ongoing cognitive, social, and physical behavior of the people who have them. That the discovered loss of a loved one typically causes grief; that grief typically causes some degree of social paralysis; these things are utterly beyond EMPATH's ken. In short, the prototypical causal roles of the several emotions are also beyond any network like EMPATH. Just as researchers have already discovered in the realm of purely physical cognition, sophisticated social cognition requires a grasp of patterns in time, and this requires that the successful network be richly endowed with recurrent

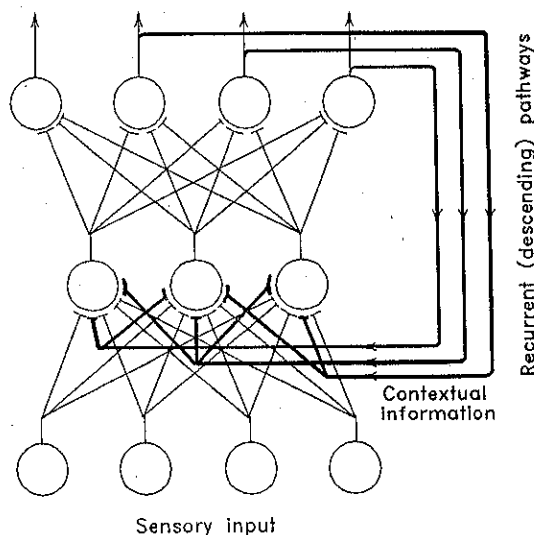


Figure 5.3  
An elementary recurrent network. The recurrent pathways are in boldface.

pathways, additional pathways that cycle information from higher neuronal layers back to earlier neuronal layers. This alone will permit the recognition of causal sequences. Figure 5.3 provides a cartoon example of a recurrent network. Such networks are trained, not on timeless snapshots, as was EMPATH. They are trained on appropriate *sequences* of input patterns.

An important subset of causal sequences is the set of ritual or conventional sequences. To take some prototypical examples, consider a social introduction, an exchange of pleasantries, an extended negotiation, a closing of a deal, a proper leave-taking. All of these mutual exchanges require, for their recognition as well as their execution, a well-tuned recurrent network. And they require of the network a considerable history spent embedded within a social space already filled with such prototypical activities on every side. After all, those prototypes must be learned, a process that will require both instructive examples and plenty of time to internalize them.

In the end, the acquired library of social prototypes hierarchically embedded in the vast neuronal activation space of any normally socialized human must rival, if it does not exceed, the acquired library of purely natural or nonsocial prototypes. One need only read a novel by someone like George Eliot or Henry James to appreciate the intricate structure of human social space and the complexity of human social dynamics. More

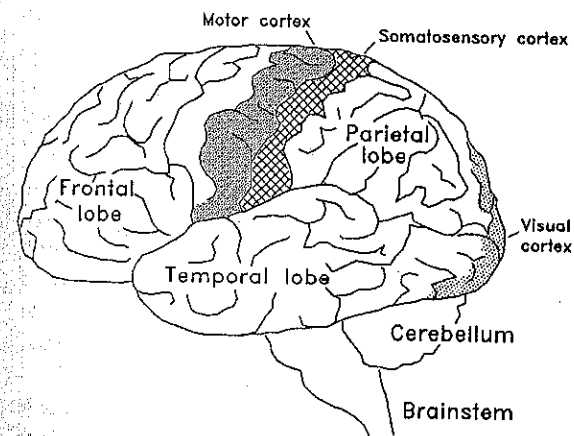


Figure 5.4  
The location of some of the primary and secondary sensory areas within the primate cerebral cortex. (Subcortical structures are not shown.) The "motor strip" or motor output cortex is also shown. Notice the broad cortical areas that lie outside these easily identified areas.

simply, recall your teenage years. Mastering that complexity is a cognitive achievement at least equal to earning a degree in physics. And yet with few exceptions, all of us do it.

#### Are There "Social Areas" in the Brain?

Experimental neuroscience in the twentieth century has focused almost exclusively on finding the neuroanatomical (i.e., structural) and the neurophysiological (i.e., activational) correlates of perceptual properties that are purely *natural* or *physical* in nature. The central and programmatic question has been as follows. Where in the brain, and by what processes, do we recognize such properties as color, shape, motion, sound, taste, aroma, temperature, texture, bodily damage, relative distance, and so on? The pursuit of such questions has led to real insights, and we have long been able to provide a map of the various areas in the brain that seem centrally involved in each of the functions mentioned.

The discovery technique is simple in concept. Insert a long, thin microelectrode into any one of the cells in the cortical area in question (the brain has no pain sensors, so the experimental animal is utterly unaware of this telephone tap), and then see whether and how that cell responds when the animal is shown color or motion, or hears tones, or feels warmth and cold, and so on. In this fashion, a functional map is painstakingly produced. Figure 5.4 provides a quick look at the several primary and secondary

sensory cortices and their positions within the rear half of a typical primate cerebral cortex.

But what about the front half of the cortex, the so-called frontal lobe? What is it for? The conventional but vague answer is, "to formulate potential motor behaviors for delivery to and execution by the motor cortex." Here we possess much less insight into the significance of these cortical structures and their neuronal activities. We cannot manipulate the input to those areas in any detail, as we can with the several sensory areas, because the input received by the premotor areas comes ultimately from all over the brain: areas that are already high up in the processing hierarchy, areas a long way from the sensory periphery where we can easily control what is and is not presented.

On the other hand, we can insert microelectrodes as before, but this time stimulate the target cell rather than record from it. In the motor cortex itself, this works beautifully. If we briefly stimulate the cells in certain areas, certain muscles in the body twitch, and there is a systematic correspondence between the areas of motor cortex and the muscles they control. In short, the motor strip itself constitutes a well-ordered map of the body's many muscles, much as the primary visual cortex is a map of the eye's retina. Stimulating single cells outside of and upstream from the motor areas, however, produces little or nothing in the way of behavioral response, presumably because the production of actual behavior requires smooth sequences of large activation vectors involving many thousands of cells at once. That kind of stimulation we still lack the technology to produce.

A conventional education in neuroscience thus leaves one wondering exactly how the entire spectrum of sensory inputs processed in the rear half of the brain finally gets transformed into some appropriate motor outputs formulated in the front half of the brain. This is indeed a genuine problem, and it is no wonder that researchers have found it so difficult. From the perspective we have gained from our study of artificial networks we can see how complex the business of vector coding and vector transformation must be in something as large as the brain.

Plainly, sleuthing out the brain's complete sensorimotor strategy would be a daunting task even if the brain were an artificial network, one whose every synaptic connection strength were known and all of whose neuronal activation levels were open to continuous and simultaneous monitoring. But a living brain is not so accommodating. Its connection strengths are mostly inaccessible, and monitoring the activity of more than a few cells at a time is currently impossible.

This is one of the reasons why the recent artificial network models have made possible so much progress. We can learn things from the models that we might never have learned from the brain directly. And we can then

return to the biological brain with some new and better-informed experimental questions to pose—questions concerning the empirical faithfulness of our network models, questions that we do have some hope of answering. Accordingly, the hidden transformations that produce behavior from perceptual input need not remain hidden after all.

If we aspire to track them down, however, we need to broaden our conception of the problem. In particular, we should be wary of the assumption that perception is first and foremost the perception of purely physical features in the world. And we should be wary of the correlative assumption that behavioral output is first and primarily the manipulation of physical objects.

We should be wary because we already know that humans and other social animals are keenly sensitive, perceptually, to *social* features of their surroundings. And because we already know that humans and social animals manipulate their *social* environment as well as their purely physical surroundings. And above all, because we already know that infants in most social species begin acquiring their *social* coordination at least as early as they begin learning sensorimotor coordination in its purely physical sense. Even infants can discriminate a smile from a scowl, a kind tone of voice from a hostile tone, a humorous exchange from a fractious one. And even an infant can successfully call for protection, induce feeding behavior, and invite affection and play.

I do not mean to suggest that social properties are anything more, ultimately, than just intricate aspects of the purely physical world. Nor do I wish to suggest that they have independent causal properties over and above what is captured by physics and chemistry. What I do wish to assert is that, in learning to represent the world, the brains of infant social creatures focus naturally and relentlessly on the social features of their local environment, often slighting physical features that will later seem unmissable. Human children, for example, typically do not acquire command of the basic color vocabulary until their third or fourth year of life, long after they have gained linguistic competence on matters such as anger, promises, friendship, ownership, and love. As a parent, I was quite surprised to discover this in my own children, and surprised again to learn that the pattern is quite general. But perhaps I should not have been. The social features listed are far more important to a young child's practical life than are the endlessly various colors.

The general lesson is plain. As social infants partition their activation spaces, the categories that form are just as often social categories as they are natural or physical categories. In apportioning neuronal resources for important cognitive tasks, the brain expends roughly as much of those resources on representing and controlling social reality as it does on representing and controlling physical reality.

Look once again, in the light of these remarks, at the brain in figure 5.3. Notice the unmapped frontal half and the large, unmapped areas of the rear half. Might some of these areas be principally involved in *social* perception and action? Might they be teeming with vast vectorial sequences representing *social* realities of one sort or other? Indeed, once the question is raised, why stop with these areas? Might the so-called primary sensory cortical areas—for touch, vision, and hearing especially—be as much in the business of grasping and processing social facts as they are in the business of grasping and processing purely physical facts? These two functions are certainly not mutually exclusive.

I think the answer is almost certainly yes to all of these questions. We lack intricate brain maps for social features comparable to existing brain maps for physical features, not because they are nonexistent, I suggest, but rather because we have not looked for them with a determination comparable to the physical case.

#### *Moral Perception and Moral Understanding*

Although there is no room to detail the case here, an examination of how neural networks sustain scientific understanding reveals that the role of learned *prototypes* and their continual redeployment in new domains of phenomena is central to the scientific process (Kuhn 1962; Churchland 1989, 1995). Specific *rules* or “laws of nature” play an undeniably important but nonetheless secondary role, mostly in the social business of communicating or teaching scientific skills. One’s scientific understanding is lodged primarily in one’s acquired hierarchy of structural and dynamical *prototypes*, not primarily in a set of linguistic formulas.

In a parallel fashion, neural network research has revealed how our knowledge of a language may be embodied in a hierarchy of *prototypes* for verbal sequences that admit of varied instances and indefinitely many combinations, rather than in a set of specific rules-to-be-followed (Elman 1992). Of course, we can and do state grammatical rules, but a child’s grammatical competence in no way depends on ever hearing them uttered or being able to state them. It may be that the main function of such rules resides in the social business of describing and refining our linguistic skills. One’s grammatical capacity, at its original core, may consist of something other than a list of internalized rules-to-be-followed.

With these two points in mind, let us finally turn to the celebrated matter of our *moral* capacity. Let us address our ability to recognize cruelty and kindness, avarice and generosity, treachery and honor, mendacity and honesty, the Cowardly Way Out and the Right Thing to Do. Here, once again, the intellectual tradition of Western moral philosophy is focused on *rules*, on specific laws or principles. These are supposed to

govern one’s behavior, to the extent that one’s behavior is moral at all. And the discussion has always centered on which rules are the truly valid, correct, or binding rules.

I have no wish whatever to minimize the importance of that ongoing moral conversation. It is an essential part of humanity’s collective cognitive adventure, and I would be honored to make even the most modest of contributions to it. Nevertheless, it may be that a normal human’s capacity for moral perception, cognition, deliberation, and action has rather less to do with rules, whether internal or external, than is commonly supposed.

What is the alternative to a rule-based account of our moral capacity? The alternative is a hierarchy of learned prototypes, for both moral perception and moral behavior, prototypes embodied in the well-tuned configuration of a neural network’s synaptic weights. We may here find a more fruitful path to understanding the nature of *moral learning*, *moral insight*, *moral disagreements*, *moral failings*, *moral pathologies*, and *moral growth* at the level of entire societies. Let us explore this alternative, just to see how some familiar territory looks from a new and different hilltop.

One of the lessons of neural network research is that one’s capacity for recognizing and discriminating perceptual properties usually outstrips one’s ability to articulate or express the basis of such discriminations in words. Tastes and colors are the leading examples, but the point quickly shows itself to have a much broader application. Faces, too, are something we can discriminate, recognize, and remember to a degree that exceeds any verbal articulation we could possibly provide. The facial expression of emotions is evidently a third example. The recognition of sounds is a fourth. In fact, the cognitive priority of the preverbal over the verbal shows itself, upon examination, to be a feature of almost all of our cognitive categories.

This supravocal grasp of the world’s many dimensions of variation is perhaps the main point of having concepts: it allows us to deal appropriately with the always novel but *never entirely* novel situations flowing endlessly toward us from an open-ended future. That same flexible readiness characterizes our social and moral concepts no less than our physical concepts. And our moral concepts show the same penetration and supravocal sophistication shown by nonmoral concepts. One’s ability to recognize instances of cruelty, patience, meanness, and courage, for instance, far outstrips one’s capacity for verbal definition of those notions. One’s diffuse expectations of their likely consequences similarly exceed any verbal formulas that one could offer or construct, and those expectations are much the more penetrating because of it. All told, moral cognition would seem to display the same profile or signature that in other domains indicates the activity of a well-tuned neural network underlying the whole process.



Figure 5.5

The old woman/young woman, a classic case of a visually ambiguous figure. The old woman is looking left and slightly toward us with her chin buried in her ruff. The young woman is looking to the left but away from us; her nose is barely visible, but her left ear, jawline, and choker necklace are directly before us.

If this is so, then moral perception will be subject to same ambiguities that characterize perception generally. Moral perception will be subject to the same modulation, shaping, and occasional "prejudice" that recurrent pathways make possible. By the same token, moral perception will occasionally be capable of the same cognitive "reversals" that we see in such examples as the old/young woman in figure 5.5. Pursuing the parallel further, it should also display cases where one's first moral reaction to a novel social situation is simply moral confusion, but where a little background knowledge or collateral information suddenly resolves that confusion into an example of something familiar, into an unexpected instance of some familiar moral prototype.

On these same assumptions, moral learning will be a matter of slowly generating a hierarchy of moral prototypes, presumably from a substantial number of relevant *examples* of the moral kinds at issue. Hence the relevance of stories and fables, and, above all, the ongoing relevance of the parental example of interpersonal behavior, and parental commentary on and consistent guidance of childhood behavior. No child can learn the route to love and laughter entirely unaided, and no child will escape the pitfalls of selfishness and chronic conflict without an environment filled with examples to the contrary.

People with moral perception will be people who have learned those lessons well. People with reliable moral perception will be those who can

protect their moral perception from the predations of self-deception and the corruptions of self-service. And, let us add, from the predations of group-think and the corruptions of fanaticism, which involves a rapacious disrespect for the moral cognition of others.

People with unusually penetrating moral insight will be those who can see a problematic moral situation in more than one way, and who can evaluate the relative accuracy and relevance of those competing interpretations. Such people will be those with unusual moral *imagination*, and a critical capacity to match. The former virtue will require a rich library of moral prototypes from which to draw, and special skills in the recurrent manipulation of one's moral perception. The latter virtue will require a keen eye for local divergences from any presumptive prototype and a willingness to take them seriously as grounds for finding some alternative understanding. Such people will by definition be rare, although all of us have some moral imagination, and all of us some capacity for criticism.

Accordingly, moral disagreements will be less a matter of interpersonal conflict over what "moral rules" to follow and more a matter of interpersonal divergence as to what moral prototype best characterizes the situation at issue. It will be more a matter, that is, of divergences over what kind of case we are confronting in the first place. Moral argument and moral persuasion, on this view, will most typically be a matter of trying to make salient this, that, or the other feature of the problematic situation, in hopes of winning one's opponent's assent to the local appropriateness of one general moral prototype over another. A nonmoral parallel of this phenomenon can again be found in the old/young woman example of figure 5.5. If that figure were a photograph, say, and if there were some issue as to what it was really a picture of, I think we would agree that the young-woman interpretation is by far the more realistic of the two. The old-woman interpretation, by comparison, asks us to believe in the reality of a hyperbolic cartoon.

A genuinely moral example of this point about the nature of moral disagreement can be found in the current issue over a woman's right to abort a first-trimester pregnancy without legal impediment. One side of the debate considers the status of the early fetus and invokes the moral prototype of a Person, albeit a very tiny and incomplete person, a person who is defenseless for that very reason. The other side of the debate addresses the same situation and invokes the prototype of a tiny and possibly unwelcome Growth, as yet no more a person than is a cyst or a cluster of one's own skin cells. The first prototype bids us bring to bear all the presumptive rights of protection due any person, especially one that is young and defenseless. The second prototype bids us leave the woman to deal with the tiny growth as she sees fit, depending on the value it may or may not currently have for her, relative to her own long-term plans as an

independently rightful human. Moral argument, in this case as elsewhere, typically consists in urging the accuracy or the poverty of the prototypes at issue as portrayals of the situation at hand.

I cite this example not to enter into this debate, nor to presume on the patience of either party. I cite it to illustrate a point about the nature of moral disagreements and the nature of moral arguments. The point is that real disagreements need not be and seldom are about what explicit moral rules are true or false. The adversaries in this case might even agree on the obvious principles lurking in the area, such as, "It is *prima facie* wrong to kill any person." The disagreement here lies at a level deeper than that glib anodyne. It lies in a disagreement about the boundaries of the category "person" and hence about whether the explicit principle even applies to the case at hand. It lies in a divergence in the way people perceive or interpret the social world they encounter and in their inevitably divergent behavioral responses to that world.

Whatever the eventual resolution of this divergence of moral cognition, it is antecedently plain that both parties to this debate are driven by some or other application of a moral prototype. But not all conflicts are thus morally grounded. Interpersonal conflicts are regularly no more principled than that between a jackal and a buzzard quarreling over a steaming carcass, or a pair of two-year-old human children screaming in frustration at a tug-of-war over the same toy. This returns us, naturally enough, to the matter of moral development in children and to the matter of the occasional failures of such development. How do such failures look, on the trained-network model here being explored?

Some of them recall a view from antiquity. Plato was inclined to argue, at least in his occasional voice as Socrates, that no man ever knowingly does wrong. For if he recognizes the action as being genuinely *wrong*—rather than just "thought to be wrong by others"—what motive could he possibly have to perform it? Generations of students have rejected Plato's suggestion, and rightly so. But Plato's point, however overstated, remains an instructive one: an enormous portion of human moral misbehavior is due primarily to *cognitive* failures of one kind or another.

Such failures are inevitable. We have neither infinite intelligence nor total information. No one is perfect. But some people, as we know, are notably less perfect than the norm, and their failures are systematic. In fact, some people are rightly judged to be chronic troublemakers, terminal narcissists, thoughtless blockheads, and treacherous snakes, not to mention bullies and sadists. Whence stem these sorry failings?

From many sources, no doubt. But we may note right at the beginning that a simple failure to develop the normal range of moral perception and social skills will account for a great deal here. Consider the child who, for whatever reasons, learns only very slowly to distinguish the minute-by-

minute flux of rights, expectations, entitlements, and duties as they are created and canceled in the course of an afternoon at the day-care center, an outing with siblings, or a playground game of hide and seek. Such a child is doomed to chronic conflict with other children—doomed to cause them disappointment, frustration, and eventually anger, all of it directed at him.

Moreover, he has all of it coming, despite the fact that a flinty-eyed determination to "flout the rules" is *not* what lies behind his unacceptable behavior. The child is a moral cretin because he has not acquired the skills already flourishing in the others. He is missing skills of recognition to begin with, and also the skills of matching his behavior to the moral circumstance at hand, even when it is dimly recognized. The child steps out of turn, seizes disallowed advantages, reacts badly to constraints binding on everyone, denies earned approval to others, and is blind to opportunities for profitable cooperation. His failure to develop and deploy a roughly normal hierarchy of social and moral prototypes may seem tragic, and it is. But one's sympathies must lie with the other children when, after due patience runs out, they drive the miscreant howling from the playground.

What holds for a playground community holds for adult communities as well. We all know adult humans whose behavior recalls to some degree the bleak portrait just outlined. They are, to put the point gently, unskilled in social practices. Moreover, all of them pay a stiff and continuing price for their failure. Overt retribution aside, they miss out on the profound and ever-compounding advantages that successful socialization brings, specifically, the intricate practical, cognitive, and emotional commerce that lifts everyone in its embrace.

#### *The Basis of Moral Character*

This quick portrait of the moral miscreant invites a correspondingly altered portrait of the morally successful person. The common picture of the Moral Man as one who has acquiesced in a set of explicit rules imposed from the outside—from God, perhaps, or from society—is dubious in the extreme. A relentless commitment to a handful of explicit rules does not make one a morally successful or a morally insightful person. That is the path of the Bible Thumper and the Waver of Mao's *Little Red Book*. The price of virtue is a good deal higher than that, and the path thereto a good deal longer. It is much more accurate to see the moral person as one who has acquired a complex set of subtle and enviable skills: perceptual, cognitive, and behavioral.

This was, of course, the view of Aristotle, to recall another name from antiquity. Moral virtue, as he saw it, was something acquired and refined

over a lifetime of social experience, not something swallowed whole from an outside authority. It was a matter of developing a set of largely inarticulable skills, a matter of *practical* wisdom. Aristotle's perspective and the neural network perspective here converge.

To see this more clearly, focus now on the single individual, one who grows up among creatures with a more or less common human nature, in an environment of ongoing social practices and presumptive moral wisdom already in place. The child's initiation into that smooth collective practice takes time—time to learn how to recognize a large variety of prototypical social situations, time to learn how to deal with those situations, time to learn how to balance or arbitrate conflicting perceptions and conflicting demands, and time to learn the sorts of patience and self-control that characterize mature skills in any domain of activity. After all, there is nothing essentially moral about learning to defer immediate gratification in favor of later or more diffuse rewards.

So far as the child's brain is concerned, such learning, such neural representation, and such deployment of those prototypical resources are all indistinguishable from their counterparts in the acquisition of skills generally. There are real successes, real failures, real confusions, and real rewards in the long-term quality of life that one's moral skills produce. As in the case of internalizing mankind's scientific knowledge, a person who internalizes mankind's moral knowledge is a more powerful, effective, and resourceful creature because of it. To draw the parallels here drawn is to emphasize the practical or pragmatic nature of both scientific and broadly normative knowledge. It is to emphasize the fact that both embody different forms of know-how: how to navigate the natural world in the former case and how to navigate the social world in the latter.

This portrait of the moral person as one who has acquired a certain family of perceptual and behavioral *skills* contrasts sharply with the more traditional accounts that picture the moral person as one who has agreed to follow a certain set of *rules* (for example, "Always keep your promises") or alternatively, as one who has a certain set of overriding *desires* (to maximize the general happiness). Both of these more traditional accounts are badly out of focus.

For one thing, it is just not possible to capture, in a set of explicit imperative sentences or rules, more than a small part of the practical wisdom possessed by a mature moral individual. It is no more possible here than in the case of any other form of expertise—scientific, athletic, technological, artistic, or political. The sheer amount of information stored in a well-trained network the size of a human brain, and the massively distributed and exquisitely context-sensitive ways in which it is stored therein, preclude its complete expression in a handful of sentences, or even a large bookful. Statable rules are not the *basis* of one's moral character.

They are merely its pale and partial reflection at the comparatively impotent level of language.

If rules do not do it, neither are suitable desires the true basis of anyone's moral character. Certainly they are not sufficient. A person might have an all-consuming desire to maximize human happiness. But if that person has no comprehension of what sorts of things genuinely serve lasting human happiness; no capacity for recognizing other people's emotions, aspirations, and current purposes; no ability to engage in smoothly cooperative undertakings; no skills whatever at pursuing that all-consuming desire; then that person is not a moral saint. He is a pathetic fool, a hopeless busybody, a loose cannon, and a serious menace to his society.

Neither are canonical desires obviously necessary. A man may have, as his most basic and overriding desire in life, the desire to see his own children mature and prosper. To him, let us suppose, everything else is distantly secondary. And yet such a person may still be numbered among the most consummately moral people of his community, so long as he pursues his personal goal, as others may pursue theirs, in a fashion that is scrupulously fair to the aspirations of others and ever protective of the practices that serve everyone's aspirations indifferently.

Attempting to portray either accepted rules or canonical desires as the basis of moral character has the further disadvantage of inviting the skeptic's hostile question: "Why should I follow those rules?" in the first case, and "What if I don't have those desires?" in the second. If, however, we reconceive strong moral character as the possession of a broad family of perceptual, cognitive, and behavioral skills in the social domain, then the skeptic's question must become, "Why should I acquire those skills?" To which the honest answer is, "Because they are easily the most important skills you will ever learn."

This novel perspective on the nature of human cognition, both scientific and moral, comes to us from two disciplines—cognitive neuroscience and connectionist artificial intelligence—that had no prior interest in or connection with either the philosophy of science or moral theory. And yet the impact on both these philosophical disciplines is destined to be revolutionary. Not because an understanding of neural networks will obviate the task of scientists or of moral-political philosophers. Not for a second. Substantive science and substantive ethics will still have to be done, by scientists and by moralists and mostly in the empirical trenches. Rather, what will change is our conception of the nature of scientific and moral knowledge, as it lives and breathes within the brains of real creatures. Thus, the impact on *metaethics* is modestly obvious already. No doubt a revolution in moral psychology will eventually have some impact on substantive ethics as well—on matters of moral training, moral pathology, and moral correction, for example. But that is for moral philosophers to work through, not cognitive

theorists. The message of this essay is that an ongoing conversation between these two communities has now become essential.

#### *Acknowledgment*

This essay is excerpted from chapters 6 and 10 of P. M. Churchland, *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*, Cambridge, Mass., 1995. I thank The MIT Press for permission to print some of that material here.

#### *References*

- Churchland, P. M. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, Mass.: Bradford Books/MIT Press.
- Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, Mass.: Bradford Books/MIT Press.
- Elman, J. L. 1992. "Grammatical Structure and Distributed Representations." In S. Davis, ed., *Connectionism: Theory and Practice*. Oxford: Oxford University Press.
- Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

## 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