Toward a Cognitive Neurobiology of the Moral Virtues

Paul M. Churchland

I. Introduction

These are the early days of what I hope will be a long and fruitful intellectual tradition, a tradition fueled by the systematic interaction and mutual information of cognitive neurobiology on the one hand and moral theory on the other. More specifically, it is the traditional sub-area we call *meta*ethics, including moral epistemology and moral psychology, that will be most dramatically informed by the unfolding developments in cognitive neurobiology. And it is metaethics again that will exert a reciprocal influence on future neurobiological research – more specifically, into the nature of moral perception, the nature of practical and social reasoning, and the development and occasional corruption of moral character.

This last point about reciprocity highlights a further point. What we are contemplating here is no imperialistic takeover of the moral by the neural. Rather, we should anticipate a mutual flowering of *both* our high-level conceptions in the domain of moral knowledge *and* our lower-level conceptions in the domain of normal and pathological neurology. For each level has much to teach the other, as this essay will try to show.

Nor need we resist this interaction of distinct traditions on grounds that it threatens to deduce normative conclusions from purely factual premises, for it threatens no such thing. To see this clearly, consider the following parallel. Cognitive neurobiology is also in the process of throwing major illumination on the philosophy of *science* – by way of revealing the several forms of neural representation that underlie scientific cognition, and the several forms of neural activity that underlie learning and conceptual change (see, for example, Churchland 1989, chapters 9–11). And yet, substantive science itself will still have to be done by scientists, according to the various methods by which we make scientific progress. An adequate theory of the brain, plainly, would not constitute a theory of Stellar Evolution or a theory of the underlying structure of the Periodic Table. It would constitute, at most, only a theory of how we generate, embody, and manipulate such worthy cognitive achievements.

Equally, and for the same reasons, substantive moral and political theory will still have to be done by moral and political thinkers, according to the various methods by which we make moral and political progress. An adequate theory of the brain, plainly, will not constitute a theory of Distributive Justice or a body of Criminal Law. It would constitute, at most, only a theory of how we generate, embody, and manipulate such worthy cognitive achievements.

These reassurances might seem to rob the contemplated program of its interest, at least to moral philosophers, but we shall quickly see that this is not the case. For we are about to contemplate a systematic and unified account, sketched in neural-network terms, of the following phenomena: moral knowledge, moral learning, moral perception, moral ambiguity, moral conflict, moral argument, moral virtues, moral character, moral pathology, moral correction, moral diversity, moral progress, moral realism, and moral unification. This collective sketch will serve at least to outline the program, and even at this early stage it will provide a platform from which to address the credentials of one prominent strand in pre-neural metaethics, the program of so-called "Virtue Ethics," as embodied in both an ancient writer (Aristotle), and three modern writers (Johnson, Flanagan, and MacIntyre).

II. The reconstruction of moral cognitive phenomena in cognitive neurobiological terms

This essay builds on work now a decade or so in place, work concerning the capacity of recent neural-network

Topoi **17**: 83–96, 1998.

models (of micro-level brain activity) to reconstruct, in an explanatory way, the salient features of molar-level cognitive activity. That research began in the mid-1980s by addressing the problems of perceptual recognition, motor-behavior generation, and other basic phenomena involving the gradual learning of sundry cognitive *skills* by artificial "neural" networks, as modeled within large digital computers (Gorman and Sejnowski, 1988; Lehky and Sejnowski, 1988; Rosenberg and Sejnowski, 1990; Lockery et al., 1990; Cottrell, 1991; Elman, 1992). From there, it has moved both downward in its focus, to try to address in more faithful detail the empirical structure of biological brains (Churchland and Sejnowski, 1992), and upward in its focus, to address the structure and dynamics of such higher-level cognitive phenomena as are displayed, for example, in the human pursuit of the various theoretical sciences (Churchland, 1989).

For philosophers, perhaps the quickest and easiest introduction to these general ideas is the highly pictorial account in Churchland (1995), to which I direct the unprepared reader. My aim here is not to recapitulate that groundwork, but to build on it. Even so, that background account will no doubt slowly emerge, from the many examples to follow, even for the reader new to these ideas, so I shall simply proceed and hope for the best.

The model here being followed is my earlier attempt to reconstruct the epistemology of the *natural* sciences in neural-network terms (Churchland, 1989). My own philosophical interests have always been centered around issues in epistemology and the philosophy of science, and so it was natural, in the mid-1980s, that I should first apply the emerging framework of cognitive neurobiology to the issues with which I was most familiar. But it soon became obvious to me that the emerging framework had an unexpected generality, and that its explanatory power, if genuine at all, would illuminate a much broader range of cognitive phenomena than had so far been addressed. I therefore proposed to extend its application into other cognitive areas such as mathematical knowledge, musical knowledge, and moral knowledge. (Some first forays appear in chapters 6 and 10 of Churchland, 1995.) These further domains of cognitive activity provide, if nothing else, a series of stiff *tests* for the assumptions and explanatory ambitions of neural-network theory. Accordingly, the present paper presumes to draw out the central theoretical claims, within the domain of metaethics, to which a neural-network model of cognition commits us. It is for the reader, and especially for professional moral philosophers themselves, to judge whether the overall portrait that results is both explanatorily instructive and faithful to moral reality.

1. *Moral knowledge*

Broadly speaking, to teach or train any neural network to embody a specific cognitive capacity is gradually to impose a specific *function* onto its input-output behavior. The network thus acquires the ability to respond, in various but systematic ways, to a wide variety of potential sensory inputs. In a simple, three-layer feedforward network with fixed synaptic connections (Figure 1a), the output behavior at the third layer of neurons is completely determined by the activity at the sensory input layer. In a (biologically more realistic) *recurrent* network (Figure 1b), the output behavior is jointly determined by sensory input *and* the prior dynamical state of the entire network. The purely feedforward case yields a cognitive capacity that is sensitive to spatial patterns but blind to temporal patterns or to temporal context; the recurrent network yields a capacity that is sensitive to, and responsive to, the changing cognitive contexts in which its sensory inputs are variously received. In both cases, the acquired cognitive capacity actually resides in the specific configuration of the many synaptic *connections* between the neuronal layers, and learning that cognitive capacity is a matter of slowly adjusting the size or "weight" of each connection so that, collectively, they come to embody the input-output function desired. On this, more in a moment.

Evidently, a trained network has acquired a specific skill. That is, it has learned how to respond, with appropriate patterns of neural activity across its output layer, to various inputs at its sensory layer. Accordingly, and as with all other kinds of knowledge, my first characterization of moral knowledge portrays it as a *set of skills*. To begin with, a morally knowledgeable adult has clearly acquired a sophisticated family of *perceptual* or *recognitional* skills, which skills allow him a running comprehension of his own social and moral circumstances, and the social and moral circumstances of the others in his community. Equally clearly, a morally knowledgeable adult has acquired a complex set of *behavioral* and *manipulational* skills, which skills make possible his successful social and moral interaction with the others in his community.

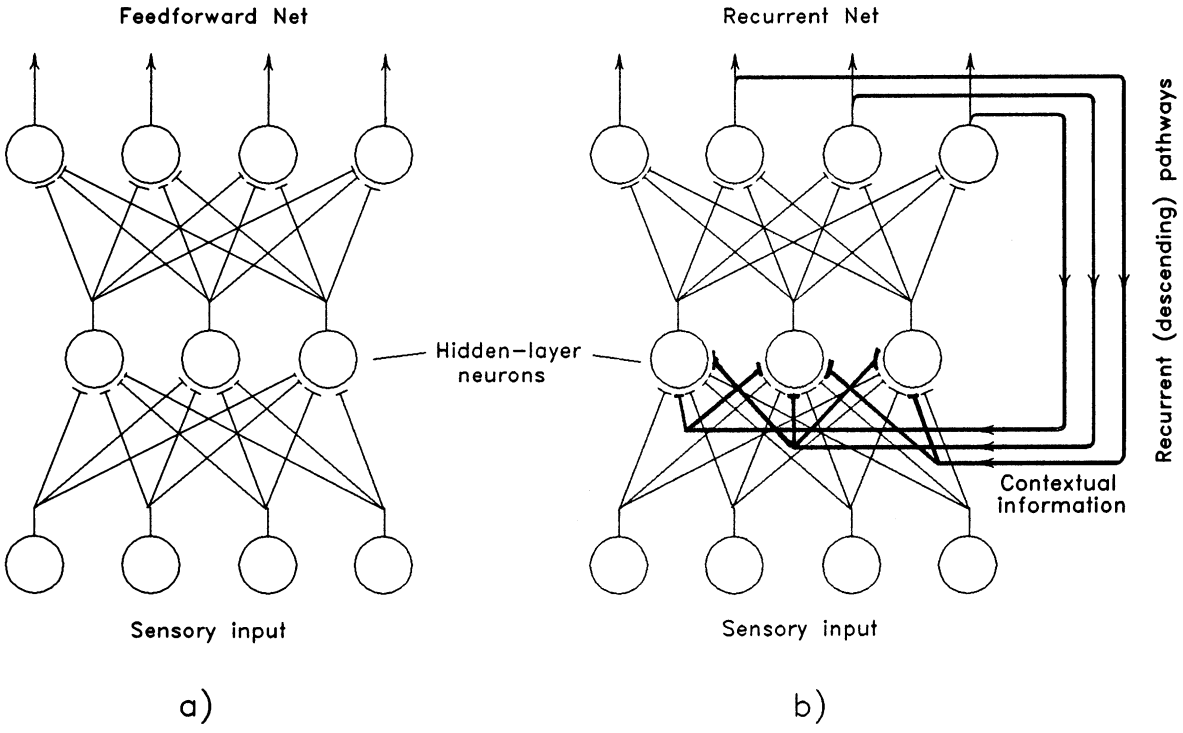

Figure 1. (a) A simple feedforward network. (b) A simple recurrent network. For a quick grip on the functional significance of such models, think of the lower or input layer of neurons as the sensory neurons, and think of the upper or output layer of neurons as the motor or muscle-driving neurons.

According to the model of cognition here being explored, the skills at issue are embodied in a vast configuration of appropriately weighted synaptic connections. To be sure, it is not intuitively obvious how a thousand, or a billion, or a trillion such connections can constitute a specific cognitive skill, but we begin to get an intuitive grasp of how they can do so when we turn our attention to the collective behavior of the neurons at the layer to which those carefully configured connections happen to attach.

Consider, for example, the second layer of the feedforward network in Figure 1a. That neuronal population, like any other discrete neuronal population, represents the various states of the world with a corresponding variety of *activation patterns* across that entire population. That is to say, just as a pattern of brightness levels across the 200,000 pixels of your familiar TV screen can represent a certain two-dimensional scene, so can the pattern of activation levels across a neuronal population represent specific aspects of the external world, although the "semantics" of that representational relation will only rarely be so obviously "pictorial." If the neuronal representation is auditory, for example, or olfactory, or gustatory, then obviously the representation will be something other than a 2-D "picture."

What is important for our purposes is that the abstract *space* of *possible* representational patterns, across a given neuronal population, slowly acquires, in the course of training the synapses, a specific structure – a structure that assigns a family of dramatically preferential abstract *locations*, within that space, in response to a preferred family of distinct stimuli at the network's sensory layer. This is how the mature network manages to categorize all possible inputs, either as rough instances of one-or-other of its learned family of prototypical *categories*, or, failing that, as instances of unintelligible noise. Before training, *all* inputs produce noise at the second layer. After training, however, that second layer has become preferentially sensitized to a comparatively tiny subset of the vast range of possible input patterns (most of which are never encountered). Those "hot-button" input patterns, whenever they occur, are subsequently assimilated to the second layer's acquired set of *prototypical categories*.

Consider an artificial network (Figure 2a) trained to discriminate human faces from nonfaces, male faces

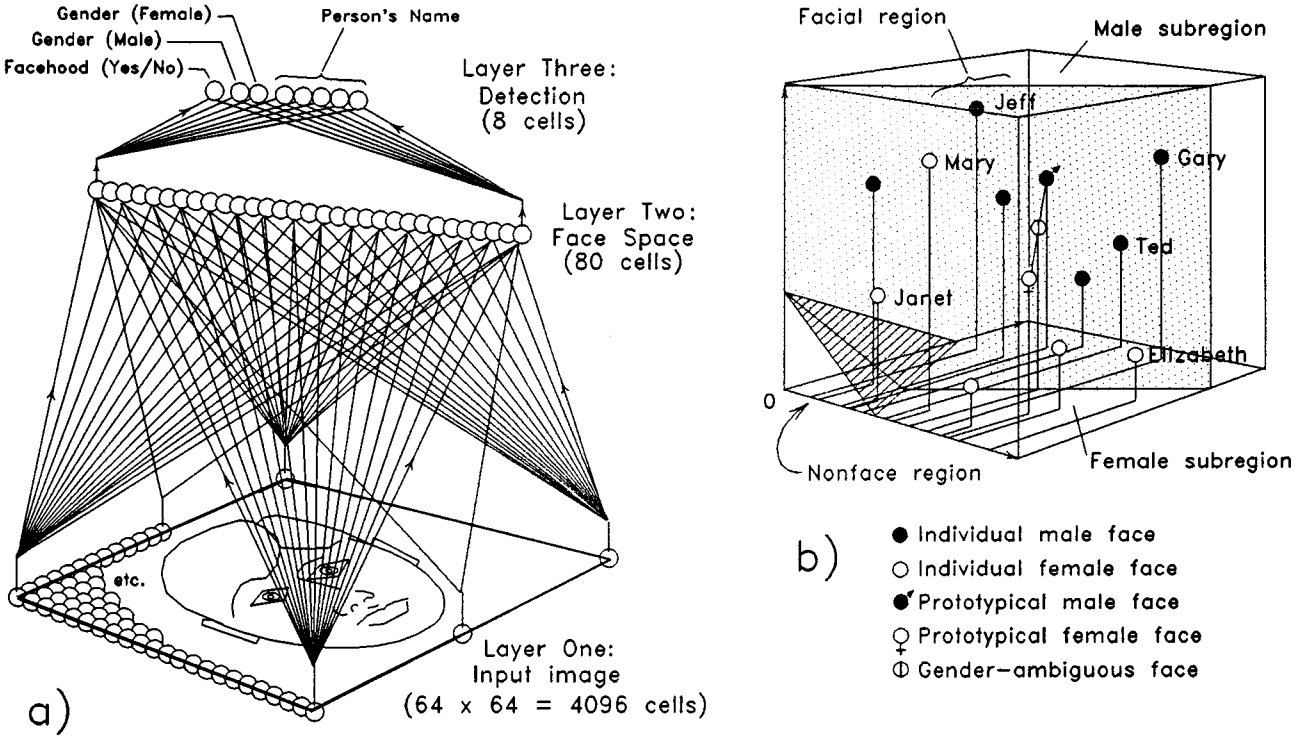

Figure 2. (a) A feedforward neural network for recognizing human faces and distinguishing gender. (b) The hierarchy of categorial partitions, acquired during training, across the space of possible neuronal activation patterns at the network's middle or "hidden" layer.

from female faces, and a handful of named individuals as presented in a variety of distinct photographs. As a result of that training, the abstract space of *possible* activation patterns across its second neuronal layer has become *partitioned* (Figure 2b), first into a pair of complementary subvolumes for neuronal activation patterns that represent sundry faces and nonfaces respectively. The former subvolume has become further partitioned into two mutually exclusive subvolumes for male faces and female spaces respectively. And within each of these two subvolumes, there are proprietary "hot-spots" for each of the named individuals that the network learned to recognize during training.

Following this simple model, the suggestion here advanced is that our capacity for *moral* discrimination also resides in an intricately configured matrix of synaptic connections, which connections also partition an abstract conceptual space, at some proprietary neuronal layer of the human brain, into a hierarchical set of categories, categories such as "morally significant" vs. "morally nonsignificant" actions; and within the former category, "morally bad" vs. "morally praiseworthy" actions; and within the former subcategory, sundry specific categories such as "lying," "cheating," "betraying," "stealing," "tormenting," "murdering," and so forth (Figure 3).

That abstract space of possible neuronal-activation patterns is a simple model for our own conceptual space for moral representation, and it displays an intricate structure of similarity and dissimilarity relations; relations that cluster similar vices close together and similar virtues close together; relations that separate highly dissimilar action categories into spatially distant sectors of the space. This high-dimensional similarity space (of course, Figure 3 ignores all but three of its many neuronal axes) displays a structured family of categorial "hot spots" or "prototype positions," to which actual sensory inputs are assimilated with varying degrees of closeness.

An abstract space of *motor*-neuron activation patterns will serve a parallel function for the generation of actual social behavior, a neuronal layer that presumably enjoys close functional connections with the sensory neurons just described. All told, these structured spaces constitute our acquired knowledge of *the structure of social space*, and *how to navigate it effectively*.

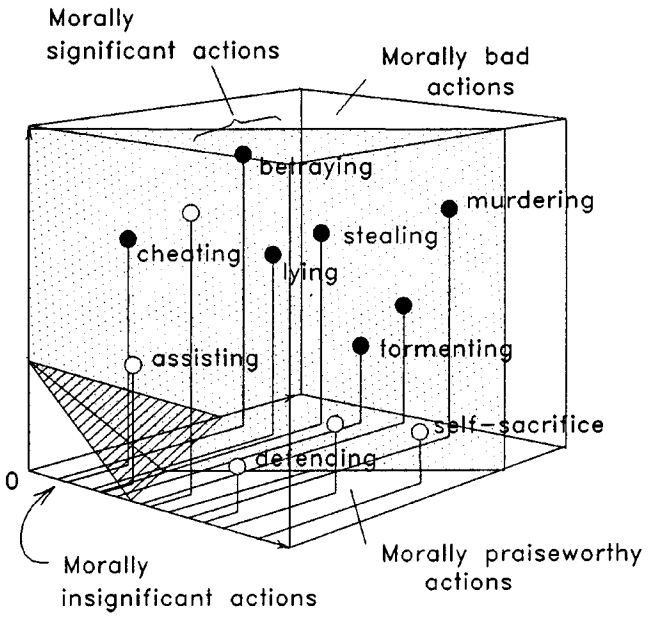

Figure 3. A (conjectural) activation space for moral discrimination.

2. *Moral learning*

Moral learning consists in the gradual generation of these internal perceptual and behavioral prototypes, a process that requires repeated exposure to, or practice of, various *examples* of the perceptual or motor categories at issue. In artificial neural networks, such learning consists in the repeated adjustment of the weights of their myriad synaptic connections, adjustments that are guided by the naive network's initial performance *failures*, as measured by a distinct "teacher" program. In living creatures, learning also consists in the repeated adjustment of one's myriad synaptic connections, a process that is also driven by one's ongoing experience with failure. Our artificial "learning technologies" are currently a poor and pale reflection of what goes on in real brains, but in both cases – the artificial networks and living brains – those gradual synaptic readjustments lead to an appropriately structured high-dimensional similarity space, a space partitioned into a hierarchical family of categorial subspaces, which subspaces contain a central hot-spot that represents a *prototypical* instance of its proprietary category.

Such learning typically takes time, often large amounts of time. And as the network models have also illustrated, such learning often needs to be structured, in the sense that the simplest of the relevant perceptual and behavioral skills need to be learned first, with the more sophisticated skills being learned later, and only after the elementary ones are in place. Moreover, such learning can display some familiar pathologies, those that derive from a narrow or otherwise skewed population of training examples. In such cases, the categorial framework duly acquired by the network fails to represent the full range and true structure of the social/moral domain it needs to represent, and performance failures are the inevitable result.

These remarks barely introduce the topic of moral learning, but we need to move on. The topic will be readdressed below, when we discuss moral progress.

3. *Moral perception*

This most fundamental of our moral skills consists in the *activation*, at some appropriate layer of neurons at least half a dozen synaptic connections away from the sensory periphery, of a specific pattern of neuronal excitation-levels that is sufficiently close to some already learned moral *prototype* pattern. That *n*th-layer activation pattern is jointly caused by the current activation pattern across one or more of the brain's sensory or input layers, and by the series of carefully trained synaptic connections that intervene. Moral perception is thus of a piece with perception generally, and its profile displays features long familiar to perceptual psychologists.

For example, one's spontaneous judgments about the social and moral configuration of one's current environment are strongly sensitive to contextual features, to collateral information, and to one's current interests and focus of attention. Moral perception is thus subject to "priming effects" and "masking effects." As well, moral perception displays the familiar tendency of cognitive creatures to "jump to conclusions" in their perceptual interpretations of partial or degraded perceptual inputs. Like artificial networks, we humans have a strong tendency automatically to assimilate our current perceptual circumstances to the nearest of the available moral prototypes that our prior training has created in us.

4. *Moral ambiguity*

A situation is morally ambiguous when it is problematic by reason of its tendency to activate *more than one*

moral prototype, prototypes that invite two incompatible or mutually exclusive subsequent courses of action. In fact, and to some degree, ambiguity is a chronic feature of our moral experience, partly because the social world is indefinitely complex and various, and partly because the interests and collateral information each of us brings to the business of interpreting the social world differ from person to person and from occasion to occasion. The recurrent or descending pathways within the brain (illustrated, in stick-figure form, in Figure 1b) provide a continuing stream of such background information (or misinformation) to the ongoing process of perceptual interpretation and prototype activation. Different "perceptual takes," on one and the same situation, are thus inevitable. Which leads us to our next topic.

5. *Moral conflict*

The activation of distinct moral prototypes can happen in two or more distinct individuals confronting the same situation, and even in a single individual, as when some contextual feature is alternatively magnified or minimized and one's overall perceptual take flips back and forth between two distinct activation patterns in the neighborhood of two distinct prototypes. In such a case, the single individual is morally conflicted ("Shall I *protect* a friend's feelings by keeping silent on someone's trivial but hurtful slur, or shall I be forthright and *truthful* in my disclosures to a friend?").

*Inter*personal conflicts about the moral status of some circumstance reflect the same sorts of divergent interpretations, driven this time by interpersonal divergences in the respective collateral information, attentional focus, hopes and fears, and other contextual elements that each perceiver brings to the ambiguous situation. Occasional moral conflicts are thus possible, indeed they are inevitable, even between individuals who had identical moral training and who share identical moral categories.

There is, finally, the extreme case where moral judgment diverges because the two conflicting individuals have fundamentally different moral conceptual frameworks, reflecting major differences in the acquired structure of their respective activation spaces. Here, even communication becomes difficult, and so does the process by which moral conflicts are typically resolved.

6. *Moral argument*

On the picture here being explored, the standard conception of moral argument as the formal deduction of moral conclusions from shared moral premises starts to look procrustean in the extreme. Instead, the administration and resolution of moral conflicts emerges as a much more dialectical process whereby the individuals in conflict take turns highlighting or making salient certain aspects of the situation at issue, and take turns urging various similarities between the situation at issue and various shared prototypes, in hopes of producing, within their adversary, an activation pattern that is closer to the prototype being defended ("It's a mindless clutch of cells, for heaven's sake! The woman is not obliged to preserve or defend it.") and/or farther from the prototype being attacked ("No, it's a miniature person! Yes, she is obliged."). It is a matter of nudging your interlocutor's current neuronal activation-point *out* of the attractor-category that has captured it, and *into* a distinct attractor-category. It is a matter of trying to change the probability, or the robustness, or the proximity to a shared neural prototype-pattern, of your opponent's neural behavior.

In the less tractable case where the opponents fail to share a common family of moral prototypes, moral argument must take a different form. I postpone discussion of this deeper form of conflict until the section on moral progress, below.

7. *Moral virtues*

These are the various skills of social *perception*, social *reflection*, *imagination*, *and reasoning*, and social *navigation* and *manipulation* that normal social learning produces. In childhood, one must come to appreciate the high-dimensional background structure of social space – its offices, its practices, its prohibitions, its commerce – and one must learn to recognize its local configuration swiftly and reliably. One must also learn to recognize one's own current position within it, and the often quite different positions of others. One must learn to anticipate the normal unfolding of this ongoing commerce, to recognize and help repair its occasional pathologies, and to navigate its fluid structure while avoiding social disasters, both large and small. All of this requires skill in divining the social perceptions and personal interests of others, and skill in manipulating and negotiating our collective behavior.

Being skills, such virtues are inevitably acquired rather slowly, as anyone who has raised children will be familiar. Nor need their continued development ever cease, at least in individuals with the continued opportunities and the intelligence necessary to refine them. The acquired structures within one's neuronal activation spaces – both perceptual and motor – can continue to be sculpted by ongoing experience and can thus pursue an ever deeper insight into, and an effectively controlling grasp of, one's enclosing social reality. Being skills, they are also differently acquired by distinct individuals, and they are differentially acquired within a single individual. Each brain is slightly different from every other in its initial physical structure, and each brain's learning history is unique in its myriad details. No two of us are identical in the profile of skills we acquire, which raises our next topic.

8. *Moral character*

A person's unique moral character is just the individual profile of his perceptual, reflective, and behavioral skills in the social domain. From what has already been said, it will be evident that moral character is distinguished more by its rich diversity across individuals than by its monotony. Given the difficulty in clearly specifying any canonical profile as being uniquely ideal, this is probably a good thing. Beyond the unending complexity of social space, the existence of a diversity of moral characters simply reflects a healthy tendency to explore that space and to explore the most effective styles of navigating it. By this, I do not mean to give comfort to moral nihilists. That would be to deny the reality of social learning. What I am underwriting here is the idea that long-term moral learning across the human race is positively served by tolerating a gaussian distribution of well-informed "experiments" rather than by insisting on a narrow and impossible orthodoxy.

This view of the assembled moral virtues as a slowly-acquired network of skills also contains an implicit critique of a popular piece of romantic nonsense, namely, the idea of the "sudden convert" to morality, as typified by the "tearful face of the repentant sinner" and the post-baptismal "born-again" charismatic Christian. Moral character is not something – is not *remotely* something – that can be acquired in a day by an Act of Will or by a single Major Insight.

The idea that it can be so acquired is a falsifying reflection of one or other of two familiar conceptions of moral character, herewith discredited. The first identifies moral character with the acceptance of a canonical set of behavior-guiding rules. The second identifies moral character with a canonical set of desires, such as the desire to maximize the general happiness, and so on. Perhaps one can embrace a set of rules in one cathartic act, and perhaps one can permanently privilege some set of desires by a major act of will. But neither act can result in what is truly needed, namely, an intricate set of finely-honed perceptual, reflective, and sociomotor skills. These take several decades to acquire. Epiphanies of moral commitment can mark, at most, the initiation of such a process. Initiations are welcome, of course, but we do not give children a high-school diploma for showing up for school on the first day of the first grade. For the same reasons, "born-again" moral characters should probably wait a similar period of time before celebrating their moral achievement or pressing their moral authority.

9. *Moral pathology*

This is a large topic, since, if there are many different ways to succeed in being a morally mature creature, there are even more ways in which one might fail. But as a first pass, moral pathology consists in the partial absence, or subsequent corruption, of the normal constellation of perceptual, reflective, and behavioral skills under discussion. In terms of the cognitive theory that underlies the present approach, it consists in the failure to achieve, or subsequently to activate normally, a suitable hierarchy of moral prototypes within one's neuronal activation space. And at the lowest level, this consists in a failure, either early or late, to achieve and maintain the proper configuration of the brain's 10^{14} synaptic weights, the configuration that sustains the desired hierarchy of prototypes and makes possible their appropriate activation.

The terms "normally," "suitable," "proper," and "appropriate" all appear in this quick characterization, and they will all owe their sense to a inextricable mix of *functional* understanding within cognitive neurobiology and genuine *moral* understanding as brought to bear by common sense and the civil and criminal law. The point here urged is that we can come to understand how displays of moral incompetence, both major and minor, are often the reflection of specific functional

failures, both large and small, within the brain. This is not a speculative remark. Thanks to the increasing availability of brain-scanning technologies such as Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI), neurologists are becoming familiar with a variety of highly specific forms of brain damage that display themselves in signature forms of cognitive failure in moral perception, moral reasoning, and social behavior (Damasio et al., 1991; Damasio, 1994; Bechara et al., 1994; Adolphs et al., 1996).

Two quick examples will illustrate the point. The neurologists Antonio and Hanna Damasio have a patient, known in the literature as "Boswell," who is independently famous for his inability to lay down any new long-term memories because of bilateral lesions to his medial temporal lobe, including his hippocampus. Since his illness, his "remembered past" is a moving window that reaches back no more than forty seconds. More importantly, for our purposes, it later emerged that he also displays a curious inability to "see evil" in pictures of various emotionally-charged and potentially violent scenes. In particular, he is unable to pick up on the various negative emotions as expressed in people's *faces*, and he will blithely confabulate innocent explanations of the socially and morally problematic scenes shown him. There is nothing wrong with Boswell's eyes or visual system, however. His cognitive deficit lies roughly a dozen synaptic steps and a dozen neuronal layers behind his retinas.

As the MRI scans revealed, Boswell's herpes-simplex encephalitis had also damaged the lower half of both of his temporal lobes, which includes the area called "IT" (infero-temporal) known for its critical role in discriminating individual human faces and in coding facial expressions. He can no longer recognize the identity of faces well-known to him before the illness (movie stars and presidents, for example), and his moral perception has been selectively impaired in the manner described.

A second patient, EVR, had a normal life as a respected accountant, devoted father and faithful husband. In his mid 40s, a ventromedial frontal brain tumor was successfully removed, and subsequent tests revealed no change in his original IQ of 140. But within six months he had lost his job for rampant irresponsibility, made a series of damaging financial decisions, was divorced by his frustrated wife, briefly married and then was left by a prostitute, and had generally become incapable of the normal prudence that guides complex planning and intricate social interactions. Subsequent MRI scans confirmed that the surgical removal of the tumor had lesioned the ventromedial frontal cortex (the seat of complex planning) and its connections to the amygdala (a primitive limbic area that apparently embodies fear and anxiety).

The functional consequence of this break in the wiring was to *isolate* EVR's practical reasonings from the "visceral" somatic and emotional reactions that normally accompany the rational evaluation of practical alternatives. In normals, those "somatic markers" (as the Damasios have dubbed them) constitute an important dimension of socially-relevant information and a key factor in inhibiting one's decisions. In EVR, they have been cut out of the loop, resulting in the sorts of behavior described above.

These two failures, of moral perception and moral behavior respectively, resulted from sudden illness and consequent damage to specific brain areas, which is what brought them to the attention of the medical profession and led to their detailed examination. But these and many other neural deficits can also appear slowly, as a result of developmental misadventures and other chronic predations – childhood infections, low-level toxins, abnormal metabolism, abnormal brain chemistry, abnormal nutrition, maternal drug use during pregnancy, and so forth. There is no suggestion, let me emphasize, that all failures of moral character can be put down to structural deficits in the brain. A proper moral education – that is, a long stretch of intricate socialization – remains a necessary condition on acquiring a well-formed moral character, and no doubt the great majority of failures, especially the minor ones, can be put down entirely to sundry inadequacies in that process.

Even so, the educational process is thoroughly entwined with the developmental process and deeply dependent on the existence of normal brain structures to embody the desired matrix of skills. At least some failures of moral character, therefore, and especially the most serious failures, are likely to involve some confounding disability or marginality at the level of brain structure and/or physiological activity. If we wish to be able wisely to address such major failures of moral character, in the law and within the correctional system, we would therefore do well to understand the many dimensions of neural failure that can collectively give rise to them. We can't fix what we don't understand.

10. *Moral correction*

Consider first the structurally and physiologically *normal* brain whose formative social environment fails to provide a normal moral education. The child's experience may lack the daily examples of normal moral behavior in others, it may lack opportunities to participate in normal social practices, it may fail to see others deal successfully and routinely with their inevitable social conflicts, and it may lack the normal background of elder sibling and parental correction of its perceptions and its behavior. For the problematic young adult that results, moral correction will obviously consist in the attempt somehow to make up a missed or substandard education.

That can be very difficult. The cognitive plasticity and eagerness to imitate found in children is much reduced in a young adult. And a young adult cannot easily find the kind of tolerant community of innocent peers and wise elders that most children are fortunate to grow up in. Thus, not one but two important windows of opportunity have been missed.

The problem is compounded by the fact that children in the impoverished social environments described do not simply fail to learn. Rather, they may learn quite well, but *what* they learn is a thoroughly twisted set of social and moral prototypes and an accompanying family of skills which – while crudely functional within the impoverished environment that shaped them, perhaps – are positively *dys*functional within the more coherent structure of society at large. This means that the young adult has some substantial *un*learning to do. Given the massive cognitive "inertia" characteristic even of normal humans, this makes the corrective slope even steeper, especially when young adult offenders are incarcerated in a closely-knit prison community of similarly twisted social agents.

This essay was not supposed to urge any substantive social or moral policies, but those who do trade in such matters may find relevant the following purely factual issues. America's budget for state and federal prisons is said to be somewhat larger than its budget for *all* of higher education, for its elite research universities, its massive state universities, its myriad liberal arts colleges, and all of its technical colleges and two-year junior colleges combined. It is at least conceivable that our enormous penal-system budget might be more wisely spent on prophylactic policies aimed at raising the quality of the social environment of disadvantaged children, rather than on policies that struggle, against much greater odds, to repair the damage once it is done.

A convulsive shift, of course, is not an option. Whatever else our prisons do or do not do, they keep at least some of the dangerously incompetent social agents and the outright predators off our streets and out of our social commerce. But the plasticity of the young over the old poses a constant invitation to shift our corrective resources childwards, as due prudence dictates. This policy suggestion hopes to reduce the absolute input to our correctional institutions. An equally important issue is how, in advance of such "utopian" advances, to increase the rate at which they are emptied, to which topic I now turn.

A final point, in this regard, about normals. The cognitive plasticity of the young – that is, their unparalled capacity for learning – is owed to neurochemical and physiological factors that fade with age. (The local production and diffusion of nitric oxide within the brain is one theory of how some synaptic connections are made selectively subject to modification, and there are others.) Suppose that we were to learn how to *recreate* in young adults, temporarily and by neuropharmacological means, that perfectly normal regime of neural plasticity and learning aptitude found in children. In conjunction with some more effective programs of resocialization than we currently command (without them, the pharmacology will be a waste of time), this might re-launch the "disadvantaged normals" into something much closer to a normal social trajectory and out of prison for good.

There remain, alas, the genuine abnormals, for whom moral correction is first a matter of trying to repair or compensate for some structural or physiological defect(s) in brain function. Even if these people are hopeless, it will serve social policy to identify them reliably, if only to keep them permanently incarcerated or otherwise out of the social mainstream. But some, at least, will not be hopeless. Where the deficit is biochemical in nature – giving rise to chronically inappropriate emotional profiles, for example – neuropharmacological intervention, in the now-familiar form of chronic subdural implants, perhaps, will return some victims to something like a normal neural economy and a normal emotional profile. That will be benefit enough, but they will then also be candidates for the resocialization techniques imagined earlier for disadvantaged normals.

This discussion presumes far more neurological understanding than we currently possess, and is plain speculative as a result. But it does serve to illustrate some directions in which we might well wish to move, once our early understanding here has matured. In any case, I shall close this discussion by reemphasizing the universal importance of gradual socialization by long interaction with a moral order already in place. We will never create moral character by medical intervention alone. There are too many trillions of synaptic connections to be appropriately weighted and only long experience can hope to do that superlatively intricate job. The whole point of exploring the technologies mentioned above will be to maximize everyone's chances of engaging in and profiting from that traditional and irreplaceable process.

11. *Moral diversity*

I here refer not to the high-dimensional bell-curve diversity of moral characters within a given culture at a given time, but to the nonidentity, across two cultures separated in space and/or in time, of the overall *system* of moral prototypes and prized skills common to most normal members of each. Such major differences in moral consciousness typically reflect differences in substantive economic circumstances between the two cultures, in the peculiar threats to social order with which they have to deal, in the technologies they command, the metaphysical beliefs they happen to hold, and other accidents of history.

Such diversity, when discovered, is often seen as grounds for a blanket scepticism about the objectivity or reality of moral knowledge. That was certainly its effect on me in my later childhood, a reaction reinforced by the astonishingly low level of moral argument I would regularly hear from my more religious schoolchums, and even from the local pulpits. But that is no longer my reaction, for throughout history there have been comparable differences, between distinct cultures, where *scientific* knowledge was concerned, and comparable block-headedness in purely "factual" reasoning (think of "New Age medicine," for example, or "UFOlogy"). But this very real diversity and equally lamentable sloppiness does not underwrite a blanket skepticism about the possibility of scientific knowledge. It merely shows that it is not easy to come by, and that its achievement requires a long-term process of careful and honest evaluation of a wide variety of complex experiments over a substantial range of human experience. Which points to our next topic.

12. *Moral progress*

If it exists – there is some dispute about this – moral progress consists in the slow change and development, over historical periods, of the moral prototypes we teach our children and forcibly impose on derelict adults, a developmental process that is gradually instructed by our collective *experience* of a collective life lived under those perception-shaping and behavior-guiding prototypes.

From the neurocomputational perspective, this process looks different only in its ontological focus – the *social* world as opposed to the *natural* world – from what we are pleased to call *scientific progress*. In the natural sciences as well, achieving adult competence is a matter of acquiring a complex family of perceptual, reflective, and behavioral skills in the relevant field. And there, too, such skills are embodied in an acquired set of structural, dynamical, and manipulational prototypes. The occasional deflationary voice to the contrary, our scientific progress over the centuries is a dramatic and encouraging reality, and it results in part from the myriad instructions (often painful) of an ongoing experimental and technological life lived under those same perception-shaping and behavior-guiding scientific prototypes.

Our conceptual development in the moral domain, I suggest, differs only in detail from our development in the scientific domain. We even have institutions whose job it is continually to fine-tune and occasionally to reshape our conceptions of proper conduct, permissible practice, and proscribed behavior. Civic, state, and federal legislative bodies spring immediately to mind, as does the civil service, and so do the several levels of the judiciary and their ever-evolving bodies of case-law and decision-guiding legal precedents. As with our institutions for empirical science, these socially-focused institutions typically outlive the people who pass through their offices, often by centuries and sometimes by many centuries. And as with the payoff from our scientific institutions, the payoff here is the accumulation of unprecedented levels of recorded (social) experience, the equilibrating benefits of collective decision making, and the resulting achievement of levels of

moral understanding that are unachievable by a single individual in a single lifetime.

To this overarching parallel it may be objected that science addresses the ultimate nature of a fixed, stable, and independent reality, while our social, legislative, and legal institutions address a plastic reality that is deeply dependent on the organizing activity of humans. But this presumptive contrast disappears almost entirely when one sees the acquisition of both scientific and moral wisdom as the acquisition of sets of *skills*. Both address a presumptively *im*plastic part of their respective domains – the basic laws of nature in the former case, and basic human nature in the latter. And both address a profoundly *plastic* part of their respective domains – the articulation, manipulation, and technological exploitation of the natural world in the case of working science, and the articulation, manipulation, and practical exploitation of human nature in the case of working morals and politics. A prosperous city represents simultaneous success in both dimensions of human cognitive activity. And the resulting artificial technologies, both natural and social, each make possible a deeper insight into the basic character of the natural universe and of human nature, respectively.

13. *Moral unity/systematicity*

This parallel with natural science has a further dimension. Just as progress in science occasionally leads to welcome unifications within our understanding – as when all planetary motions come to be seen as special cases of projectile motion, and all optical phenomena come to be seen as special cases of electromagnetic waves – so also does progress in moral theory bring occasional attempts at conceptual unification – as when our assembled obligations and prohibitions are all presented (by Hobbes) as elements of a *social contract*, or (by Kant) as the local instantiations of a *categorical imperative*, or (by Rawls) as the reflection of *rules rationally chosen from behind a veil of personal ignorance*. These familiar suggestions, and others, are competing attempts to unify and systematize our scattered moral intuitions or antecedent moral understanding, and they bring with them (or hope to bring with them) the same sorts of virtues displayed by intertheoretic reductions in science, namely, greater simplicity in our assembled conceptions, greater consistency in their application, and an enhanced capacity (born of increased generality) for dealing with novel kinds of social and moral problems.

As with earlier aspects of moral cognition, this sort of large-scale cognitive achievement is also comprehensible in neurocomputational terms, and it seems to involve the very same sorts of neurodynamical changes that are (presumptively) involved when theoretical insights occur within the natural sciences. Specifically, a wide range of perceptual phenomena – which (let us suppose) used to activate a large handful of distinct moral prototypes, m_1, m_2, m_3, . . . , m_n – come to be processed under a new regime of recurrent manipulation (recall the recurrent neuronal pathways of Figure 1b) that results in them all activating an unexpected moral prototype *M*, a prototype whose typical deployment has hitherto been in other perceptual domains entirely, a prototype that now emerges as a *superordinate* prototype of which the scattered lesser prototypes, m_1, m_2, m_3, . . . , m_n can now be seen, retrospectively, as so many *sub*ordinate instances.

The preceding is a neural-network description of what happens when, for example, our scattered knowledge in some area gets *axiomatized*. But axiomatization, in the linguaformal guise typically displayed in textbooks, is but one minor instance of this much more general process, a process that embraces the many forms of *non*discursive knowledge as well, a process that embraces science and ethics alike.

14. *Reflections on some recent "virtue ethics"*

As most philosophers will perceive, the general portrait of moral knowledge that emerges from neural-network models of cognition is a portrait already under active examination within moral philosophy, quite independently of any connections it might have with cognitive neurobiology. Its original champion is Aristotle and its current research community includes figures as intellectually diverse as Mark Johnson (1993), Owen Flanagan (1991), and Alasdair MacIntyre (1981), all of whom came to this general perspective for reasons entirely of their own. For the many reasons outlined in the body of this paper, I am compelled (and honored) to count myself among them. But I am not entirely comfortable in this group, for two of the philosophers just mentioned take a view, on the matter of moral progress, very different from that just outlined. Flanagan (1996) has expressed frank doubts that human moral con-

sciousness ever makes much genuine "progress," and he suggests that its occasional changes are better seen as just a directionless meander made in local response to our changing economic and social environment.

MacIntyre (1981) voices a different but comparably skeptical view, wherein he hankers after the lost innocence of pre-Enlightenment human communities, which were much more tightly knit by a close fabric of shared social practices, which practices provided the sort of highly interactive and mutually-dependent environment needed for the many moral virtues to develop and flourish. He positively laments the emergence of the post-Enlightenment, liberal, secular, and comparatively anonymous and independent social lives led by modern industrial humans, since the rich soil necessary for moral learning, he says, has thereby been impoverished. The familiar moral virtues must now be acquired, polished, and exercised in what is, comparatively, a social vacuum. If anything, in the last few centuries we have suffered a moral *re*gress.

I disagree with both authors, and will close by outlining why. I begin with MacIntyre, and I begin by conceding his critique of the (British) Enlightenment's cartoonlike conception of *homo economicus*, a hedonic calculator almost completely free of any interest in or resources for evaluating the very desires that drive his calculations. I likewise concede his critique of the (Continental) Enlightenment's conception of *pure reason* as the key to identifying a unique set of behavior-guiding rules. And my concessions here are not reluctant. I agree wholeheartedly with MacIntyre that neither conception throws much light on the nature of moral virtue.

But as crude as these moral or meta-moral ideas were, they were still a step up from the even more cartoonlike conceptions of *homo sheepicus* and *homo infanticus* relentlessly advanced by the pre-Enlightenment Christian Church. Portraying humanity as sheep guided by a supernatural Shepherd, or as children beholden to a supernatural Father, was an even darker self-deception and was even less likely to serve as a means by which to climb the ladder of moral understanding.

I could be wrong in this blunt assessment, and if I am, so be it. For the claim of the preceding paragraph does *not* embody the truly important argument for moral progress at the hands of the Enlightenment. That argument lies elsewhere. It lies in the permanent opening of a tradition of cautious *tolerance* for a diversity of local communities each bonded by their own fabric of social practices; it lies in the establishment of lasting institutions for the principled *evaluation* of diverse modes of social organization, and for the institutionalized *criticism* of some and the systematic *emulation* of others. It lies, in sum, in the fact that the Enlightenment broke the hold of a calcified moral dictatorship and replaced it with a tradition that was finally prepared to learn from its deliberately broad experience and its inevitable mistakes in first-order moral policy.

Once again, I am appealing to a salient parallel. The virtue of the Enlightenment, in the moral sphere, was precisely the same virtue displayed in the scientific sphere, namely, the legitimation of responsible theoretical diversity and the establishment of lasting institutions for its critical evaluation and positive exploitation. It is this long-term process, rather than any particular moral theory or moral practice that might fleetingly engage its attention, that marks the primary achievement of the Enlightenment.

MacIntyre began his Introduction to *After Virtue* with a thought-provoking science fiction scenario about the loss of an intricate practical tradition that alone gives life to its corresponding family of theoretical terms, and the relative barrenness of their continued use in the absence of that sustaining tradition. This embodies the essentials of his critique of our moral history since the Enlightenment. But we can easily construct, for critical evaluation, a parallel critique of our *scientific* history since the same period, and that parallel, I suggest, throws some welcome light on MacIntyre's rather conservative perspective.

Consider the heyday of Aristotelian Science, from the fourth century B.C. to the seventeenth century A.D. (even longer than the Christian domination of the moral sphere), and consider the close-knit and unifying set of intellectual and technological practices that it sustained. There is the medical tradition running from Rome's Galen to the four Humors of the late medieval doctors. There is the astronomical/astrological tradition that extends through Alexandria's Ptolemy to Prague's Johannes Kepler, who was still casting horoscopes for the wealthy despite his apostate theorizing. There is the intricate set of industrial practices maintained by the alchemists from the Alexandrian Greeks to seventeenth-century Europe, which tradition simply owned the vital practices of metallurgy and metal-working, and of dye-making and medicinal manufacture as well. These three traditions, and others that space bids me pass over, were

closely linked by daily practice as well as by conceptual ancestry, and they formed a consistent and coherent environment in which the practical and technological virtues of late antiquity could flourish. As they did. MacIntyre's first condition is met.

So is his second, for this close-knit "paradise" is well and truly lost, having been displaced by a hornet's nest of distinct sciences, sciences as diverse as astrophysics, molecular biology, anthropology, electrical engineering, solid-state physics, immunology, and thermodynamic meteorology. Modern science now addresses and advances on so many fronts that the research practice of individual scientists and the technological practice of individual engineers is increasingly isolated from all but the most immediate members of their local cognitive communities. And the cognitive virtues they display are similarly fragmented. They may even find it difficult to talk to each other.

You see where I am going. There may well be problems – real problems – arising from the unprecedented flourishing of the many modern sciences, but losing an earlier and somehow more healthy "golden age" is certainly not one of them. Though real, those problems are simply the price that humanity pays for growing up, and we already attempt to address them by way of interdisciplinary curricula, interdisciplinary conferences and anthologies, and by the never-ending search for explanatory unifications and intertheoretic reductions.

I propose, for MacIntyre's reflection, a parallel claim for our moral, political, and legal institutions since the Enlightenment. Undoubtedly there are problems emerging from the unprecedented flourishing of the many modern industrial societies and their sub-societies, but losing touch with a prior golden age is not obviously one of them. The very real problems posed by moral and political diversity are simply the price that humanity pays for growing up. And as in the case of the scattered sciences, we already attempt to address them by constant legislative tinkering, by the reality-driven evolution of precedents in the judicial record, by tolerating the occasional political "divorce" (e.g., Yugoslavia, the Soviet Union, the Scottish Parliament), and by the never-ending search for legal, political, and economic unifications. Next to the discovery of Fire and the polydoctrinal example of ancient Greece, the Enlightenment may be the best thing that ever happened to us.

The doctrinal analog of MacIntyre's implicit Communitarianism in moral theory is a hyperbolic form of Kuhnian conservatism in the philosophy of science, a conservativism that values the (very real) virtues of any given "normal science" tradition (such as Ptolemaic astronomy, classical thermodynamics, or Newtonian mechanics) over the comparatively fragile institutions of collective evaluation, comparison, and criticism that might slowly force their hidden vices into the sunlight and pave the way for their rightful overthrow at the hands of even more promising modes of cognitive organization. One can certainly see Kuhn's basic "communitarian" point: stable scientific practices make many valuable things possible. But tolerant institutions for the evaluation and modification of those practices make even *more* valuable things possible – most obviously, new and more stable practices.

This particular defense of the Enlightenment also lays the foundation for my response to Flanagan's quite different form of skepticism. As I view matters from the neural-network perspective explained earlier in this essay, I can find no difference in the presumptive brain mechanisms and cognitive processes that underwite moral cognition and scientific cognition. Nor can I find any significant differences in the respective social institutions that administer our unfolding scientific and moral consciousness respectively. In both cases, learning from experience is the perfectly normal outcome of both the neural and the social machinery. That means that moral progress is no less possible and no less likely than scientific progress. And since none of us, at this moment, is being shown the instruments of torture in the Vatican's basement, I suggest it is actual as well.

There remains the residual issue of whether the *sciences* make genuine progress, but that issue I leave for another time. The take-home claims of the present essay are that, 1) whatever their ultimate status, moral and scientific cognition are on an *equal* footing, since they use the same neural mechanisms, show the same dynamical profile, and respond in both the short and the long term to similar empirical pressures; and 2) in both moral and scientific learning, the fundamental cognitive achievement is the acquisition of *skills*, as embodied in the finely-tuned configuration of the brain's 10^{14} synaptic connections.

References

Adolphs, R., Tranel, D., Bechara, A., Damasio, H., and Damasio, A. R.: 1996, 'Neuropsychological Approaches to Reasoning and Decision Making', in A. R. Damasio et al. (Eds.), *The Neurobiology of Decision-Making*, Berlin: Springer-Verlag, pp. 157–180.

Bechara, A., Damasio, A., Damasio, H., and Anderson, S. W.: 1994, 'Insensitivity to Future Consequences Following Damage to Human Prefrontal Cortex', *Cognition* **50**, 7–15.

Churchland, P. M.: 1989, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge: The MIT Press.

Churchland, P. M.: 1989a, 'On the Nature of Theories: A Neurocomputational Perspective', in W. Savage (Ed.), *Scientific Theories: Minnesota Studies in the Philosophy of Science*, Vol. XIV, Minneapolis: University of Minnesota Press, pp. 59–101. Chapter 9 of Churchland, P. M. 1989.

Churchland, P. M.: 1989b, 'On the Nature of Explanation: A PDP Approach', Chapter 10 of Churchland, P. M. 1989. Reprinted in J. Misiek (Ed.), *Rationality*, Vol. 175 of *Boston Studies in the Philosophy of Science*, Dordrecht: Kluwer, 1995.

Churchland, P. M.: 1989c, 'Learning and Conceptual Change', Chapter 11 of Churchland, P. M. 1989.

Churchland, P. M.: 1995, *The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain*, Cambridge: The MIT Press.

Cottrell, G.: 1991, 'Extracting Features from Faces Using Compression Networks: Face, Identity, Emotions and Gender Recognition Using Holons', in D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton (Eds.), *Connectionist Models: Proceedings of the 1990 Summer School*, Morgan Kaufmann: San Mateo, CA.

Damasio, A. R.: 1994, *Descartes' Error*, New York: Putnam & Sons.

Damasio, A. R., Tranel, D., and Damasio, H.: 1991. 'Somatic Markers and the Guidance of Behavior', in H. Levin et al. (Eds.), *Frontal Lobe Function and Dysfunction*, New York: Oxford University Press.

Elman, J. L.: 1992, 'Grammatical Structure and Distributed Representations', in S. Davis (Ed.), *Connectionism: Theory and Practice*, Vol. 3 in the series Vancouver Studies in Cognitive Science. Oxford: Oxford University Press.

Flanagan, O.: 1991, *Varieties of Moral Personality: Ethics and Psychological Realism*, Cambridge: Harvard University Press.

Flanagan, O.: 1996, 'The Moral Network', in B. McCauley (Ed.), *The Churchlands and Their Critics*, Cambridge, Mass.: Blackwell, pp. 192–215.

Gorman, R. P. and Sejnowski, T. J.: 1988, 'Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets', *Neural Networks* **1**, 75–89.

Johnson, M.: 1993, *Moral Imagination*, Chicago: Chicago University Press.

Lehky, S. and Sejnowski, T. J.: 1988, 'Network Model of Shape-from-Shading: Neuronal Function Arises from Both Receptive and Projective Fields', *Nature* **333**, 452–454.

Lehky, S. and Sejnowski, T. J.: 1990, 'Neural Network Model of Visual Cortex for Determing Surface Curvature from Images of Shaded Surfaces', *Proceedings of the Royal Society of London* **B240**, 251–278.

Lockery, S. R., Fang, Y., and Sejnowski, T. J.: 1991, 'A Dynamical Neural Network Model of Sensorimotor Transformation in the Leech', *Neural Computation* **2**, 274–282.

MacIntyre, A.: 1981, *After Virtue*, Notre Dame: University of Notre Dame Press.

Rosenberg, C. R. and Sejnowski, T. J.: 1987, 'Parallel Networks that Learn to Pronounce English Text', *Complex Systems* **1**, 145–168.

Saver, J. L. and Damasio, A. R.: 1991, 'Preserved Access and Processing of Social Knowledge in a Patient with Acquired Sociopathy due to Ventromedial Frontal Damage', *Neuro-psychologia* **29**, 1241–1249.

Department of Philosophy
UCSD