Paul Churchland

Neurophilosophy *at* Work

CAMBRIDGE www.cambridge.org/9780521864725

This page intentionally left blank

Neurophilosophy at Work

In this collection of essays, Paul Churchland explores the unfolding impact of the several empirical sciences of the mind, especially cognitive neurobiology and computational neuroscience, on a variety of traditional issues central to the discipline of philosophy. Representing Churchland's most recent investigations, they continue his research program, launched more than thirty years ago, which has evolved into the field of neurophilosophy. Topics such as the nature of consciousness, the nature of cognition and intelligence, the nature of moral knowledge and moral reasoning, neurosemantics or "world representation" in the brain, the nature of our subjective sensory qualia and their relation to objective science, and the future of philosophy itself are here addressed in a lively, graphical, and accessible manner. Throughout the volume, Churchland's view that science is as important as philosophy is emphasized. Several of the colored figures in the volume will allow readers to perform some novel phenomenological experiments on their own visual system.

Paul Churchland holds the Valtz Chair of Philosophy at the University of California, San Diego. One of the most distinguished philosophers at work today, he has received fellowships from the Andrew Mellon Foundation, the Woodrow Wilson Center, the Canada Council, and the Institute for Advanced Study in Princeton. A former president of the American Philosophical Association (Pacific Division), he is the editor and author of many articles and books, most recently *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain* and *On the Contrary: Critical Essays, 1987–1997* (with Patricia Churchland).

Neurophilosophy at Work

PAUL CHURCHLAND

University of California, San Diego

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521864725

First published in print format 2007

ISBN-13 978-0-511-27163-2 eBook (Adobe Reader)
ISBN-10 0-511-27163-8 eBook (Adobe Reader)

ISBN-13 978-0-521-86472-5 hardback
ISBN-10 0-521-86472-0 hardback

ISBN-13 978-0-521-69200-7 paperback
ISBN-10 0-521-69200-8 paperback

Contents

Preface

Any research program is rightly evaluated on its unfolding ability to address, to illuminate, and to solve a broad range of problems antecedently recognized by the professional community. The research program at issue in this volume is cognitive neurobiology, a broad-front scientific research program with potential relevance to a considerable variety of intellectual disciplines, including neuroanatomy, neurophysiology, neurochemistry, neuropathology, developmental neurobiology, psychiatry, psychology, artificial intelligence, and . . . philosophy. It is the antecedently recognized problems of this latter discipline in particular that constitute the explanatory challenges addressed in the present volume. My aim in what follows is to direct the light of computational neuroscience and cognitive neurobiology – or such light as they currently provide – onto a range of familiar philosophical problems, problems independently at the focus of much fevered philosophical attention.

Some of those focal problems go back at least to Plato, as illustrated in Chapter 8, where we confront the issue of how the mind grasps the timeless structure underlying the ephemeral phenomena of the perceivable world. And some go back at least to Aristotle, as illustrated in Chapters 3 and 4, where we confront the issue of how the mind embodies and deploys the moral wisdom that slowly develops during the social maturation of normal humans. Other problems have moved into the spotlight of professional attention only recently, as in Chapter 1, where we address the ground or nature of consciousness. Or as in Chapter 7, where we address the prospects of artificial intelligence. Or as in Chapter 9, where we confront the allegedly intractable problems posed by subjective sensory qualia. But all of these problems look interestingly different when viewed from the perspective of recent developments in the empirical/theoretical research program of cognitive neurobiology. The low-dimensional 'box canyons', in which conventional philosophical approaches have become trapped, turn out to be embedded within higher dimensions of doctrinal possibility, dimensions in which specific directions of development appear both possible and promising. Once we have freed ourselves from the idea that cognition is basically a matter of manipulating sentence-like states (the various 'propositional attitudes' such as perceives-that-*P*, believes-that-*P*, suspects-that-*P*, and so on), according to rules of deductive and inductive inference, and once we have grasped the alternative modes of world representation, information coding, and information processing displayed in all terrestrial brains, each of the problems listed earlier appears newly tractable and potentially solvable.

The distributed illumination here promised is additionally intriguing because it comes from a single source – the vector-coding and vector/matrix-processing account of the brain's cognitive activity – an empirically based account of how the brain represents the world, and of how it manipulates those representations. Such a 'consilience of inductions', as William Whewell would describe it, lends further credence to the integrity of the several solutions proposed. The solutions proposed are not 'independent' solutions: they will stand, or fall, together.

As the reader will discover, all but one of the essays here collected were written in response, either explicit or implicit, to the published researches of many of my distinguished academic colleagues,[1] and each embodies my attempts to exploit, expand, and extend the most noteworthy contributions of those colleagues, and (less often, but still occasionally) to resist, reconstruct, or subvert them. Though cognitive neurobiology hovers always in the near background, the overall result is less a concerted argument for a specific thesis, as in a standard monograph, but more a many-sided conversation in a parlor full of creative and resourceful interlocutors. To be sure, my voice will dominate the pages to follow, for these are my essays. But the voices of my colleagues will come through loud and clear even so, partly because of their intrinsic virtues, and partly because the point of these essays is to try to address and answer those voices, not to

[1] The exception is Chapter 5, the essay on American educational policy, specifically, on the antiscience initiatives recently imposed, and since rescinded, in Kansas. I had thought these issues to be safely behind us, but after the 2004 elections, fundamentalist initiatives are once again springing up all over rural America, including, once again, poor Kansas. The lessons of this particular essay are thus newly germane.

muffle them. Without those voices, there would have been no challenges to answer, and no essays to collect.

The result is also a journey through a considerable diversity of philosophical subdisciplines, for the voices here addressed are all in hot pursuit of diverse philosophical enthusiasms. In what follows, we shall explore contemporary issues in the nature of consciousness itself, the fortunes of nonreductive materialism (specifically, functionalism) in the philosophy of mind, the neuronal basis of our moral knowledge, the future of our moral consciousness, the roles of science and religion in our public schools, the proper cognitive kinematics for the epistemology of the twenty-first century, the basic nature of intelligence, the proper semantic theory for the representational states of terrestrial brains generally, the fortunes of scientific realism, recent arguments against the identity theory of the mind–brain relation, the fundamental differences between digital computers and biological brains, the neuronal basis of our subjective color qualia, the existence of novel – indeed, 'impossible' – color qualia, and the resurrection of objective colors from mere 'secondary' properties to real and important features of physical surfaces. What unites these scattered concerns is, once more, that they are all addressed from the standpoint of the emerging discipline of cognitive neurobiology. The exercise, as a whole, is thus a test of that discipline's systematic relevance to a broad spectrum of traditional philosophical issues. Whether, and how well, it passes this test is a matter for the reader to judge. My hopes, as always, are high, but the issue is now in your hands.

Provenances

"Catching Consciousness in a Recurrent Net," first appeared in A. Brook and D. Ross, eds., *Daniel Dennett: Contemporary Philosophy in Focus,* pp. 64–81 (Cambridge: Cambridge University Press, 2002).

"Functionalism at Forty: A Critical Retrospective," first appeared in *Journal of Philosophy* 102, no. 1 (2005): 33–50.

"Toward a Cognitive Neurobiology of the Moral Virtues," first appeared in *Topoi* 17 (1998): 1–14, a special issue on moral reasoning.

"Rules, Know-How, and the Future of Moral Cognition," first appeared in *Moral Epistemology Naturalized,* R. Campbell and B. Hunter, eds., *Canadian Journal of Philosophy*, suppl. vol. 26 (2000): 291–306.

"Science, Religion, and American Educational Policy," first appeared in *Public Affairs Quarterly* 14, no. 4 (2001): 279–91.

"What Happens to Reliabilism When It Is Liberated from the Propositional Attitudes?" first appeared in *Philosophical Topics,* 29, no. 1 and 2 (2001): 91–112, a special issue on the philosophy of Alvin Goldman.

"On the Nature of Intelligence: Turing, Church, von Neumann, and the Brain," first appeared in S. Epstein, ed., *A Turing-Test Sourcebook*, ch. 5 (The MIT Press 2006).

"Neurosemantics: On the Mapping of Minds and the Portrayal of Worlds," first appeared in K. E. White, ed., *The Emergence of Mind,* pp. 117–47 (Milan: Fondazione Carlo Elba, 2001).

"Chimerical Colors: Some Phenomenological Predictions from Cognitive Neuroscience," first appeared in *Philosophical Psychology* 18, no. 5 (2005).

"On the Reality (and Diversity) of Objective Colors: How Color-Qualia Space Is a Map of Reflectance-Profile Space," is currently in press at *Philosophy of Science* (2006).

"Into the Brain: Where Philosophy Should Go from Here," first appeared in *Topoi* 25 (2006): 29–32, a special issue on the future of philosophy.

1

Catching Consciousness in a Recurrent Net

Dan Dennett is a closet Hegelian. I say this not in criticism, but in praise, and hereby own to the same affliction. More specifically, Dennett is convinced that human cognitive life is the scene or arena of a swiftly unfolding evolutionary process, an essentially cultural process above and distinct from the familiar and much slower process of biological evolution. This superadded Hegelian adventure is a matter of a certain style of *conceptual* activity; it involves an endless contest between an evergreen variety of conceptual *alternatives*; and it displays, at least occasionally, a welcome *progress* in our conceptual sophistication, and in the social and technological practices that structure our lives.

With all of this, I agree, and will attempt to prove my fealty in due course. But my immediate focus is the peculiar *use* to which Dennett has tried to put his background Hegelianism in his provocative 1991 book, *Consciousness Explained.*[1] Specifically, I wish to address his peculiar account of the *kinematics and dynamics* of the Hegelian Unfolding that we both acknowledge. And I wish to query his novel *deployment* of that kinematics and dynamics in explanation of the focal phenomenon of his book: consciousness. To state my negative position immediately,

[1] (Boston: Little, Brown, 1991). I first addressed Dennett's account of consiousness in *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain* (Cambridge, MA: MIT Press, 1995), 264–9. A subsequent two-paper symposium appears as S. Densmore and D. Dennett, "The Virtues of Virtual Machines," and P. M. Churchland, "Densmore and Dennett on Virtual Machines and Consciousness," *Philosophy and Phenomenological Research* 59, no. 3 (Sept., 1999): 747–67. This essay is my most recent contribution to our ongoing debate, but Dennett has a worthy reply to it in a recent collection of essays edited by B. L. Keeley, *Paul Churchland* (New York: Cambridge University Press, 2005), 193–209.

I am unconvinced by his declared account of the background process of human conceptual evolution and development – specifically, the Dawkinsean account of rough gene-analogs called "memes" competing for dominance of human cognitive activity.[2] And I am even less convinced by Dennett's attempt to capture the emergence of a peculiarly human consciousness in terms of our brains' having internalized a specific complex *example* of such a "meme," namely, the serial, discursive style of cognitive processing typically displayed in a von Neumann computing machine.

My opening task, then, is critical. I think Dennett is wrong to see human consciousness as the result of a unique form of "software" that began running on the existing hardware of human brains some ten, or fifty, or a hundred thousand years ago. He is importantly wrong about the character of that background software process in the first place, and he is wrong again to see consciousness itself as the isolated result of its "installation" in the human brain. Instead, as I shall argue, the phenomenon of consciousness is the result of the brain's basic *hardware* structures, structures that are widely shared throughout the animal kingdom, structures that produce consciousness in meme-free and von Neumann–innocent animals just as surely and just as vividly as they produce consciousness in us. As my title indicates, I think the key to understanding the peculiar weave of cognitive phenomena gathered under the term "consciousness" lies in understanding the dynamical properties of biological neural networks with a highly *recurrent* physical architecture – an architecture that represents a widely shared hardware feature of animal brains generally, rather than a unique software feature of human brains in particular.

On the other hand, Dennett and I share membership in a small minority of theorists on the topic of consciousness, a small minority even among materialists. Specifically, we both seek an explanation of consciousness in the *dynamical* signature of a conscious creature's cognitive activities rather than in the peculiar character or subject matter of the *contents* of that creature's cognitive states. Dennett may seek it in the dynamical features of a "virtual" von Neumann machine, and I may seek it in the dynamical features of a massively recurrent neural network, but we are both working the "dynamical profile" side of the street, in substantial isolation from the rest of the profession.

Accordingly, in the second half of this paper I intend to defend Dennett in this dynamical tilt, and to criticize the more popular content-focused

[2] As outlined in M. S. Dawkins, *The Selfish Gene* (Oxford: Oxford University Press, 1976), and Dawkins,*The Extended Phenotype* (San Francisco: Freeman, 1982).

alternative accounts of consciousness, as advanced by most philosophers and even by some neuroscientists. And in the end, I hope to convince both Dennett and the reader that the hardware-focused recurrent-network story offers the most fertile and welcoming reductive home for the relatively unusual dynamical-profile approach to consciousness that Dennett and I share.

I. Epistemology: Naturalized and Evolutionary

Attempts to reconstruct the canonical problems of epistemology within an explicitly evolutionary framework have a long and vigorous history. Restricting ourselves to the twentieth century, we find, in 1934, Karl Popper already touting experimental falsification as the selectionist mechanism within his expressly evolutionary account of scientific growth, an account articulated in several subsequent books and papers.[3] In 1950, Jean Piaget published a broader and much more naturalistic vision of information-bearing structures in a three-volume work assimilating biological and intellectual evolution.[4] Thomas Kuhn's 1962 classic[5] painted an overtly antilogicist and anticonvergent portrait of our scientific development, and proposed instead a radiative process by which different cognitive paradigms would evolve toward successful domination of a wide variety of cognitive niches. In 1970, and partly in response to Kuhn, Imre Lakatos[6] published a generally Popperian but much more detailed account of the dynamics of intellectual evolution, one more faithful to the logical, sociological, and historical facts of our own scientific history. In 1972, Stephen Toulmin[7] was pushing a biologized version of Hegel, and by 1974 Donald Campbell[8] had articulated a deliberately Darwinian account of the blind generation and selective retention of scientific theories over historical time.

[3] *Logik der Forschung* (Wien, 1934). Published in English as *The Logic of Scientific Discovery* (London: Hutchison, 1980). See also Poppers's *locus classicus* essay, "Conjectures and Refutations," in his *Conjectures and Refutations* (London: Routledge, 1972). See also Popper, *Objective Knowledge: An Evolutionary Approach* (Oxford: Oxford University Press, 1979).

[4] *Introduction a l'epistemologie genetique*, 3 vols. (Paris: Presses Universitaires de France, 1950). See also Piaget, *Insights and Illusions of Philosophy* (New York: Meridian Books, 1965), and Piaget, *Genetic Epistemology* (New York: Columbia University Press 1970).

[5] *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962).

[6] "Falsification and the Methodology of Scientific Research Programs," in I. Lakatos and A. Musgrave, eds., *Criticism and the Growth of Knowledge* (Cambridge: Cambridge University Press, 1970).

[7] S. Toulmin, *Human Understanding* (Princeton, NJ: Princeton University Press, 1972).

[8] "Evolutionary Epistemology," in *The Philosophy of Karl Popper*, P. A. Schilpp, ed. (La Salle, IL: The Open Court, 1974).

From 1975 on, the literature becomes too voluminous to summarize easily, but it includes Richard Dawkins's specific views on memes, as scouted briefly in *The Selfish Gene* (1976) and more extensively in *The Extended Phenotype* (1982). In some respects, Dawkins's peculiar take on human intellectual history is decidedly better than the take of many others in this tradition – most important, his feel for both genetic theory and biological reality is much better than that of his precursors. In other respects, it is rather poorer – comparatively speaking, and once again by the standards of the tradition at issue. Dawkins is an epistemological naïf, and his feel for our actual scientific/conceptual history is rudimentary. But he had the wit, over most of his colleagues, to escape the biologically naïve construal of theories-as-*genotypes* or theories-as-*phenotypes* that attracted so many other writers. Despite a superficial appeal, both of these analogies are deeply strained and ultimately infertile, both as extensions of existing biological theory and as explanatory contributions to existing epistemological theory.[9] Dawkins embraces, instead, and despite my opening characterization, a theories-as-*viruses* analogy, wherein the human brain serves as a host for competing invaders, invaders that can replicate by subsequently invading as-yet uninfected brains.

While an improvement in several respects, this analogy seems stretched and problematic still, at least to these eyes. An individual virus is an individual physical thing, locatable in space and time. An individual theory is no such thing. And even its individual "tokens" – as they may be severally embodied in the distinct brains they have "invaded" – are, at best, abstract *patterns* of some kind imposed upon preexisting physical structures within the brain, not physical *things* bent on making further physical things with a common physical structure.

Further, a theory has no internal mechanism that effects a literal self-replication when it finds itself in a fertile environment, as a virus has when it injects its own genetic material into the interior of a successfully hijacked cell. And my complaint here is not that the mechanisms of self-replication are different across the two cases. It is that there *is no* such mechanism for theory tokens. If they can be seen as "replicating" at all, it must be by some wholly different process. This is further reflected in the fact that theory tokens do not replicate themselves *within* a given individual, as viruses most famously do. For example, you might have 10^6

[9] An insightful perspective on the relevant defects is found in C. A. Hooker, *Reason, Regulation, and Realism: Toward a Regulatory Systems Theory of Reason and Evolutionary Epistemology* (Albany, NY: SUNY Press, 1995), 36–42.

qualitatively identical rhinoviruses in your system at one time, all children of an original invader; but never more than one token of Einstein's theory of gravity.

Moreover, the brain is a medium selected precisely for its ability to assume, hold, and deploy the conceptual systems we call theories. Theories are not alien invaders bent on subverting the brain's resources to their own selfish "purposes." On the contrary, a theory is the brain's way of making sense of the world in which it lives, an activity that is its original and primary function. A bodily cell, by contrast, enjoys no such intimate relationship with the viruses that intrude upon its normal metabolic and reproductive activities. A mature cell that is completely free of viruses is just a normal, functioning cell. A mature brain that is completely free of theories or conceptual frameworks is an utterly dysfunctional system, barely a brain at all.

Furthermore, theories often – indeed, usually – take *years* of hard work and practice to grasp and internalize, precisely because there is no analog to the physical virus entering the body, pill-like or bullet-like, at a specific time and place. Instead, a vast reconfiguration of the brain's 10^{14} synaptic connections is necessary in order to imprint the relevant conceptual framework on the brain, a reconfiguration that often takes months or years to complete. Accordingly, the "replication story" needed, on the Dawkinsean view, must be nothing short of an entire theory of how the brain *learns*. No simple "cookie-cutter" story of replication will do for the dubious "replicants" at this abstract level. There are no zipper-like molecules to divide down the middle and then reconstitute themselves into two identical copies. Nor will literally repeating the theory, by voice or in print, to another human do the trick. Simply receiving, or even memorizing, a list of presented *sentences* (a statement of the theory) is not remotely adequate to successful acquisition of the conceptual framework to be replicated, as any unprepared student of classical physics learns when he or she desperately confronts the problem-set on the final examination, armed only with a crib sheet containing flawless copies of Newton's gravitation law and the three laws of motion. Knowing a theory is not just having a few lines of easily transferable syntax, as the student's inevitable failing grade attests.

The poverty of its "biological" credentials aside, the *explanatory payoff* for embracing this viruslike conception of theories is quite unremarkable in any case. The view brings with it no compelling account of where theories originate, how they are modified over time in response to experimental evidence, how competing theories are evaluated, how they guide our experimental and practical behaviors, how they fuel our technological economies, and how they count as representations of the world's hidden structure. In short, the analogy with viruses does not provide particularly illuminating answers, or any answers at all, to most of the questions that make up the problem-domain of epistemology and the philosophy of science.

What it does do is hold out the promise of a grand consilience – a conception of scientific activity that is folded into a larger and more powerful background conception of biological processes in general. This is, at least in prospect, an extremely *good* thing, and it more than accounts for the "aha!" feelings that most of us experience upon first contemplating such a view. But closer examination shows it to be a *false* consilience, based on a false analogy. Accordingly, we should not have much confidence in deploying it, as Dennett does, in hopes of illuminating either human cognitive development in general, or the development of human consciousness in particular.

Despite reaching a strictly negative conclusion here, not just about the theories-as-viruses analogy but about the entire evolutionary tradition in recent epistemology, I must add that I still regard that tradition as healthy, welcome, and salutary, for it seeks a worthy sort of consilience, and it serves as a vital foil against the deeply sclerotic logicist tradition of the logical empiricists. Moreover, I share the background conviction of most people working in the newer tradition – namely, that in the end a proper account of human scientific knowledge must somehow be a proper part of a general theory of biological systems and biological development. However, I have quite different expectations about how that integration should proceed. They are the focus of a book in progress, but the present occasion is focused on consciousness, so I must leave their articulation for another time. In the meantime, I recommend C. A. Hooker's "nested hierarchy of regulatory mechanisms" attempt – to locate scientific activity within the embrace of biological phenomena at large – as the most promising account in the literature.[10] We now return to Dennett.

II. The Brain as Host for the von Neumann Meme

If the human brain *were* a von Neumann machine (hereafter, vN machine) – literally, rather than figuratively or virtually – then the virus

[10] Hooker, *Reason, Regulation, and Realism,* 36–42. For a review of Hooker's book and its positive thesis, see P. M. Churchland, "Review of *Reason, Regulation, and Realism,*" *Philosophy and Phenomenological Research* 58, no. 4 (1999): 541–4.

analogy just rejected would have substantially more point. We do speak of, and bend resources to avoid, "computer viruses," and the objections voiced earlier, concerning theories and the brain, are mostly irrelevant if the virus analogy is directed instead at programs loaded in a computer. A program *is* just a package of syntax; a program *can* download in seconds; a program *can* contain a self-copying subroutine; and a program *can* fill a hard drive with monotonous copies of itself, whether or not it ever succeeds in infecting a second machine.

But the brains of animals and humans are most emphatically *not* vN machines. Their coding is not digital; their processing is not serial; they do not execute stored programs; and they have no random-access storage registers whatever. As fifty years of neuroscience and fifteen years of neuromodeling have taught us, a brain is a different kettle of fish entirely. That is why brains are so hopeless at certain tasks, such as multiplying two twenty-digit numbers in one's head, which task a computer does in a second. And that is why computers are so hopeless at certain other tasks, such as recognizing individual faces or understanding speech, which task a brain does in even less time.

We now know enough about both brains and vN computers to appreciate precisely why the brain does as well as it does, despite being made of components that are a million times slower than those of an electronic computer. Specifically, the brain is a massively parallel vector processor. Its background understanding of the world's general features (its conceptual framework) resides in the slowly acquired configuration of its 10^{14} synaptic connections. Its specific understanding of the local world here-and-now (its fleeting thoughts and perceptions) resides in the fleeting patterns or vectors of activation-levels across its 10^{11} neurons. And the character of those fleeting patterns is dictated by the learned matrix of synaptic connections that serve simultaneously to transform *peripheral* sensory activation vectors into well-informed *central* vectors, and ultimately into the well-orchestrated *motor* vectors that produce our bodily behavior.

Now Dennett knows all of this as well as anyone, and it poses a problem for him. It's a problem because, as discussed earlier, the virus analogy that he intends to exploit requires a vN computer for its plausibility. But the biological brain is not a vN computer. So Dennett postulates that, at some point in our past, the human brain managed to "reprogram" itself in such a fashion that its genetically endowed "hardware" came to "load" and "run" a peculiar piece of novel "software" – an invading virus or meme – such that the brain came to *be* a "virtual" von Neumann machine.

But wait a minute. We are here contemplating an explanation – of how the brain *came to be* a virtual vN machine – in terms that make clear and literal sense only if the brain was *already* a (literal) vN machine. But it wasn't. And so it couldn't become *any* new "virtual" machine – and a fortiori not a virtual vN machine – in the literal fashion described. Dennett must have some related but metaphorical use in mind for the expressions "program," "software," "hardware," "load," and "run." And, as we shall see, for "virtual" and "vN machine" as well.

As indeed he does. Dennett knows that brains are plastic in their configurations of synaptic connections, and he knows that changing those configurations produces changes in the way the brain processes information. He is postulating that, at some point in the past, at least one human brain lucked/stumbled into a global configuration of synaptic connections that embodied an importantly new style of information processing, a style that involved, at least occasionally, the sequential, temporally structured, rule-respecting kinds of activities seen in a typical vN machine.

Let us look into this possibility. What is the actual potential of a massively parallel vector-processing machine to "simulate" a vN machine? For a purely feedforward network (Figure 1.1*a*), it is zero, because such a network cannot execute the temporally *recursive* procedures essential to a program-executing vN machine. To surmount this trivial limitation, we need to step up to networks with a *recurrent* architecture (Figure 1.1*b*), for as is well known, this is what permits any neural network to deal with structures in time.

Artificial recurrent networks have indeed been trained up to execute successfully the kinds of explicitly recursive procedures involved in, for example, adding individual pairs of *n*-digit numbers,[11] and distinguishing grammatical from ungrammatical sentences in a (highly simplified) productive language.[12]

But are these suitably trained networks thus "virtual" adders and "virtual" parsers? No. They are *literal* adders and parsers. The language of "virtual machines" is not strictly appropriate here, because these are *not* cases of a special purpose "software machine" running, qua program, on a vN-style universal Turing machine.

More generally, the idea that a machine, any machine, might be programmed to "simulate" a vN machine in particular makes the mistake of treating *vN machine* as if it were itself a *special*-purpose piece of software,

[11] G. W. Cottrell, and F. Tsung, "Learning Simple Arithmetic Procedures," *Connection Science* 5, no. 1 (1993): 37–58.

[12] J. L. Elman, "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, 1992), 138–94.

rather than what it is, namely, an entirely *general*-purpose organization of *hardware*. In sum, the brain is not a machine that is capable of "downloading software" in the first place, and a vN machine is not a piece of "software" fit for downloading in any case.

Accordingly, I cannot find a workable interpretation of Dennett's proposal here that is both nonmetaphorical and true. Dennett seems to be trying to both eat his cake (the brain becomes a vN machine by downloading some software) and have it too (the brain is not a vN machine to begin with). And these complaints are additional to and independent of the complaints of the preceding section, to the effect that Dawkins's virus analogy for cultural acquisitions such as theories, songs, and practices is a false and explanatorily sterile analogy to begin with.

There is an irony here. The fact is, if we do look to recurrent neural networks – which brains most assuredly are – in order to purchase something like the functional properties of a vN machine, we no longer *need* to "download" any epigenetically supplied meme or program, because the sheer hardware configuration of a recurrent network already delivers the desired capacity for recognizing, manipulating, and generating serial structures in time, right out of the box. Those characteristic recurrent pathways are the very computational resource that allows us to recognize a puppy's gait, a familiar tune, a complex sentence, and a mathematical proof. Which *particular* temporal structures come to dominate a network's cognitive life will be a function of which causal processes are perceptually encountered during its learning phase. But the need for a virtual vN machine, in order to achieve this broader family of cognitive ends, has now been lifted. The brain doesn't need to import the "software" Dennett contrives for it: its existing "hardware" is already equal to the cognitive tasks that he (rightly) deems important.

This fact moves me to try to reconstruct a vaguely Dennettian account of consciousness using the very real resources of a recurrent physical architecture, rather than the strained and figurative resources of a virtual vN machine. And this brings me to the dynamical-profile approach cited at the outset of this paper. But first I must motivate its pursuit by evoking and dismantling its principal explanatory adversary, the content-focused approach.

III. Consciousness as Self-Representation: Some Problems

One strategy for trying to understand consciousness is to see it as a species of *representation*, a species distinguished by its peculiar *contents*, specifically, the current states or activities of the *self*, that is, the current states or activities of the very biological-cum-cognitive system engaged in such representation. Consciousness, on this view, is essentially a matter of self-perception or self-representation. Thus, one is conscious when, for example, one's cognitive system represents stress or damage to some part of one's body (pain), when it represents one's empty stomach (hunger), when it represents the postural configuration of one's body (hands folded in front of one), when it represents one's high-level cognitive state ("I believe Budapest is in Hungary"), or when it represents one's relation to an external object ("I'm about to be hit by an incoming snowball").

Kant's doctrine of inner sense in *The Critique of Pure Reason* is the classic (and highly a priori) instance of this approach, and Antonio Damasio's book *The Feeling of What Happens*[13] provides a modern (and neurologically grounded) instance of the same general strategy. While I have some sympathy for this approach to consciousness – I have defended it myself in *Matter and Consciousness*[14] – this chapter is aimed at overturning it and replacing it with a specific alternative. Let me begin by voicing the central worries – to which all parties must be sensitive – that cloud the self-representation approach to consciousness.

There are two major weaknesses in the approach. The first is that it fails, at least on all outstanding versions, to give a clear and adequate account of the inescapable distinction between those of our self-representations that are conscious and those that are not. The nervous system has a great many subsystems that continuously monitor a wide variety of visceral, hormonal, thermal, metabolic, and other regulatory activities of the biological organism. These are representations of the self, if anything is, but they are only occasionally a part of our consciousness, and some of them are *permanently* beneath the level of conscious awareness.

One might try to avoid this difficulty by stipulating that the self-representations that constitute the domain of consciousness must be representations of the states and activities of the brain and nervous system proper, rather than of the body in general. But this proposal has three daughter difficulties. Prima facie, the stipulation would *exclude* far too much, for hunger, pain, and other plainly conscious somatosensory sensations are clearly representations of various aspects of the body, not the brain. Less obviously, but equally problematic, it would falsely *include* the

[13] (New York: Harcourt,1999).

[14] Rev. ed. (Cambridge, MA: MIT Press, 1986), 73–5, 119–20, 179–80.

enormous variety of brain activities that constitute ongoing and systematic representations of other aspects of the brain itself – indeed, these are the bulk of them – but which never make it into the spotlight of consciousness. We must be mindful, that is, that most of the brain's representational activities are self-directed and lie well below the level of conscious awareness. Finally, the proposed stipulation would wrongly *exclude* from consciousness the brain's unfolding representations of the world beyond the body, such as our visual awareness of the objects at arm's length and our auditory awareness of the whistling kettle. One might try to insist that, strictly speaking, it is only our visual and auditory *sensations* of which we are directly conscious – external objects being only indirect and secondary objects of awareness – but this move is false to the facts of both human cognitive development and human phenomenology, and it leads us down the path of classical sense-datum theory, whose barrenness has long been apparent.

A special *subject matter*, then, seems not to be the essential feature that distinguishes conscious representations from all others. To the contrary, it would seem that a conscious representation could have any content or subject matter at all. The proposal under discussion would seem to be confusing *self*-consciousness with consciousness in general. The former is highly interesting, to be sure, but it is the latter that is our current explanatory target.

The self-representation view has a second major failing, which emerges as follows. Consider a creature, such as you or me, who has a battery of distinct sensory modalities – a visual system, an auditory system, an olfactory system – for constructing representations of various aspects of the physical world. And suppose further that, as cognitive theorists, we have some substantial understanding of how those several modalities actually work, as devices for monitoring aspects of external reality and coding those aspects internally. And yet we remain mystified about what makes the representations in which they trade *conscious* representations. We remain mystified, that is, at what distinguishes the conscious states of one's visual system from the equally representational but utterly unconscious representational states of a voltmeter, an audio tape recorder, or a video camera. Now, if our general problem is thus to try to understand how *any* representational modality ascends to the level of conscious representations, then proposing a proprietary representational modality whose job it is to monitor phenomena *inside* the skin, rather than outside the skin, is a blatant case of *repeating* our problem, not of solving it. Our original problem attends the inward-looking modality no less than the various outward-looking modalities with which we began, and adding the inward modality does nothing obvious to transform the outward ones in any case. Once again, leaning on the *content* of the representations at issue – on the *focus, target,* or *subject matter* of the epistemic modality in question – fails to provide the explanatory factors that we seek. We need to look elsewhere.

IV. The Dynamical-Profile Approach

We need to look, I suggest, at the peculiar *activities* in which some of our representations participate, and at the special computational context required for those activities to take place. I here advert, for example, to the brain's capacity (1) to focus attention on some aspect or subset of its teeming polymodal sensory inputs, (2) to try out different conceptual interpretations of that selected subset, and (3) to hold the results of that selective/interpretive activity in short-term memory for long enough (4) to update a coherent representational "narrative" of the world-unfolding-in-time, a narrative thus fit for possible selection and imprinting in long-term memory.

Any cognitive representation that figures in the dynamical/computational profile just outlined is a recognizable candidate for, and a presumptive instance of, the class of *conscious* representations. We may wish to demand still more of such candidates than merely meeting these quick four conditions, but even these four specify a dynamical or functional profile recognizable as typical of conscious representations. Notice also that this profile makes no reference to the specific *content,* either semantic or qualitative, of the representation that meets it, reflecting the fact, agreed to in the last section, that a conscious representation could have any content whatever.

Appealing to notions such as attention, interpretation, and short-term memory may seem, however, to be just helping oneself to a handful of notions that are as opaque or problematic as the notion of consciousness itself, unless we can provide independent explanations of these dynamical notions in neuronal terms. In fact, that is precisely what the dynamical properties of recurrent neural networks allow us to do, and more besides, as I shall now try to show.

The consensus concerning information processing in artificial neural networks is that their training history slowly produces a *sculpted space* of possible representations (= possible activation patterns) at any given layer or population of neurons (such as the middle layer of the network in Figure 1.1*a*). Such networks, trained to discriminate or recognize

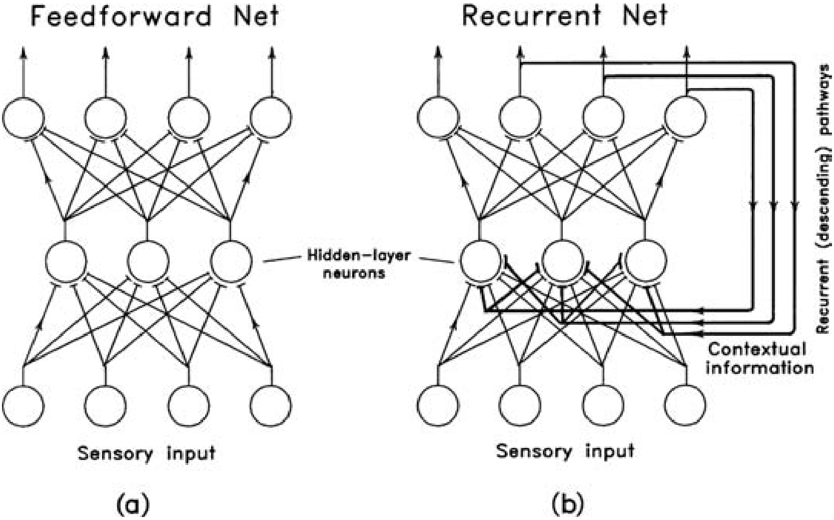

FIGURE 1.1. Elementary networks

instances of some range of categories, $c_1, \ldots, c_2$, slowly acquire a corresponding family of “attractors” or “prototype wells” variously located within the space of possible activation patterns. That sculpted space *is* the conceptual framework of that layer of neurons. Diverse sensory-layer instances of those learned perceptual categories produce activation patterns within, or close to, one or another of these “preferred” prototype regions within the activation space of the second layer of neurons.

Purely feedforward networks can achieve quite astonishing levels of discriminatory skill, but beyond a welcome tendency to “fill in” or “complete” degraded or partial perceptual instances of the categories to which they have been trained,[15] they are rather dull and predictable fellows. However, if we add recurrent or descending pathways to the basic feedforward architecture, as in Figure 1.1*b*, we lift ourselves into a new universe of functional and dynamical possibilities.

For example, information from the higher levels of any network – information that is the result of somewhat earlier information processing by the network – can be entered as a supplementary “context fixer” at the second layer of the network. This information can and does serve to “prime” or “prejudice” that neuronal population’s collective activity in the direction of one or another of its learned perceptual categories.

[15] See pp. 45–6 and 107–14 of Churchland, *The Engine of Reason, the Seat of the Soul*, for a more detailed discussion of this intriguing feature of feedforward network activity.