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.