Patricia Smith Churchland

How do neurons know?

My knowing *anything* depends on my neurons – the cells of my brain.[1] More precisely, what I know depends on the specific configuration of connections among my trillion neurons, on the neurochemical interactions between connected neurons, and on the response portfolio of different neuron types. All this is what makes me *me*.

The range of things I know is as diverse as the range of stuff at a yard sale. Some is knowledge how, some knowledge that, some a bit of both, and some not exactly either. Some is fleeting, some enduring. Some I can articulate, such as the instructions for changing a tire, some, such as how I construct a logical argument, I cannot.

Some learning is conscious, some not. To learn some things, such as how to ride a bicycle, I have to try over and over; by contrast, learning to avoid eating oysters if they made me vomit the last time just happens. Knowing how to change a tire depends on cultural artifacts, but knowing how to clap does not.

And *neurons* are at the bottom of it all. How did it come to pass that we know *anything*?

Early in the history of living things, evolution stumbled upon the advantages accruing to animals whose nervous systems could make predictions based upon past correlations. Unlike plants, who have to take what comes, animals are movers, and having a brain that can learn confers a competitive advantage in finding food, mates, and shelter and in avoiding dangers. Nervous systems earn their keep in the service of prediction, and, to that end, map the *me-relevant* parts of the world – its spatial relations, social relations, dangers, and so on. And, of course, brains map their worlds in varying degrees of complexity, and relative to the needs, equipment, and lifestyle of the organisms they inhabit.[2]

Patricia Smith Churchland is UC President's Professor of Philosophy and chair of the philosophy department at the University of California, San Diego, and adjunct professor at the Salk Institute. She is past president of the American Philosophical Association and the Society for Philosophy and Psychology. Her latest books are "Brain-Wise: Studies in Neurophilosophy" (2002) and "On the Contrary: Critical Essays, 1987 – 1997" (with Paul Churchland, 1998).

1 Portions of this paper are drawn from my book *Brain-Wise: Studies in Neurophilosophy* (Cambridge, Mass.: MIT Press, 2002).

2 See Patricia Smith Churchland and Paul M. Churchland, "Neural Worlds and Real Worlds," *Nature Reviews Neuroscience* 3 (11) (November 2002): 903 – 907.

Thus humans, dogs, and frogs will represent the same pond quite differently. The human, for example, may be interested in the pond's water source, the potability of the water, or the potential for irrigation. The dog may be interested in a cool swim and a good drink, and the frog, in a good place to lay eggs, find flies, bask in the sun, or hide.

Boiled down to essentials, the main problems for the neuroscience of knowledge are these: How do structural arrangements in neural tissue embody knowledge (the problem of representations)? How, as a result of the animal's experience, do neurons undergo changes in their structural features such that these changes constitute knowing something new (the problem of learning)? How is the genome organized so that the nervous system it builds is able to learn what it needs to learn?

The spectacular progress, during the last three or four decades, in genetics, psychology, neuroethology, neuroembryology, and neurobiology has given the problems of how brains represent and learn and get built an entirely new look. In the process, many revered paradigms have taken a pounding. From the ashes of the old verities is arising a very different framework for thinking about ourselves and how our brains make sense of the world.

Historically, philosophers have debated how much of what we know is based on instinct, and how much on experience. At one extreme, the rationalists argued that essentially all knowledge was innate. At the other, radical empiricists, impressed by infant modifiability and by the impact of culture, argued that all knowledge was acquired.

Knowledge displayed at birth is obviously likely to be innate. A normal neonate rat scrambles to the warmest place, latches its mouth onto a nipple, and begins to suck. A kitten thrown into the air rights itself and lands on its feet. A human neonate will imitate a facial expression, such as an outstuck tongue. But other knowledge, such as how to weave or make fire, is obviously learned postnatally.

Such contrasts have seemed to imply that everything we know is either caused by genes or caused by experience, where these categories are construed as exclusive and exhaustive. But recent discoveries in molecular biology, neuroembryology, and neurobiology have demolished this sharp distinction between nature and nurture. One such discovery is that normal development, right from the earliest stages, relies on both genes and epigenetic conditions. For example, a female (XX) fetus developing in a uterine environment that is unusually high in androgens may be born with male-looking genitalia and may have a masculinized area in the hypothalamus, a sexually dimorphic brain region. In mice, the gender of adjacent siblings on the placental fetus line in the uterus will affect such things as the male/female ratio of a given mouse's subsequent offspring, and even the longevity of those offspring.

On the other hand, paradigmatic instances of long-term learning, such as memorizing a route through a forest, rely on genes to produce changes in cells that embody that learning. If you experience a new kind of sensorimotor event during the day – say, for example, you learn to cast a fishing line – and your brain rehearses that event during your deep sleep cycle, then the gene *zif*-268 will be up-regulated. Improvement in casting the next day will depend on the resulting gene products and their role in neuronal function.

Indeed, five important and related discoveries have made it increasingly clear

just how interrelated 'nature' and 'nurture' are, and, consequently, how inadequate the old distinction is.[3]

First, what genes do is code for proteins. Strictly speaking, there is no gene for a sucking reflex, let alone for female coyness or Scottish thriftiness or cognizance of the concept of zero. A gene is simply a sequence of base pairs containing the information that allows RNA to string together a sequence of amino acids to constitute a protein. (This gene is said to be 'expressed' when it is transcribed into RNA products, some of which, in turn, are translated into proteins.)

Second, natural selection cannot directly select particular wiring to support a particular domain of knowledge. Blind luck aside, what determines whether the animal survives is its behavior; its equipment, neural and otherwise, underpins that behavior. Representational prowess in a nervous system can be selected for, albeit indirectly, only if the representational package informing the behavior was what gave the animal the competitive edge. Hence representational sophistication and its wiring infrastructure can be selected for only via the behavior they upgrade.

Third, there is a truly stunning degree of conservation in structures and developmental organization across all vertebrate animals, and a very high degree of conservation in basic cellular functions across phyla, from worms to spiders to humans. All nervous systems use essentially the same neurochemicals, and their neurons work in essentially the same way, the variations being vastly outweighed by the similarities. Humans have only about thirty thousand genes, and we differ from mice in only about three hundred of those;[4] meanwhile, we share about 99.7 percent of our genes with chimpanzees. Our brains and those of other primates have the same organization, the same gross structures in roughly the same proportions, the same neuron types, and, so far as we know, much the same developmental schedule and patterns of connectivity.

Fourth, given the high degree of conservation, whence the diversity of multicellular organisms? Molecular biologists have discovered that some genes regulate the expression of other genes, and are themselves regulated by yet other genes, in an intricate, interactive, and systematic organization. But genes (via RNA) make proteins, so the expression of one gene by another may be affected via sensitivity to protein products. Additionally, proteins, both within cells and in the extracellular space, may interact with each other to yield further contingencies that can figure in an unfolding regulatory cascade. Small differences in regulatory genes can have large and far-reaching effects, owing to the intricate hierarchy of regulatory linkages between them. The emergence of complex, interactive cause-effect profiles for gene expression begets very fancy regulatory cascades that can beget very fancy organisms – us, for example.

Fifth, various aspects of the development of an organism from fertilized egg to up-and-running critter depend on where and when cells are born. Neurons originate from the daughter cells of the last division of pre-neuron cells. Whether such a daughter cell becomes a glial (supporting) cell or a neuron, and which type of some hundred types of neurons

3 In this discussion, I am greatly indebted to Barbara Finlay, Richard Darlington, and Nicholas Nicastro, "Developmental Structure in Brain Evolution," *Behavioral and Brain Sciences* 24 (2) (April 2001): 263 – 278.

4 See John Gerhart and Marc Kirschner, *Cells, Embryos, and Evolution* (Oxford: Blackwell, 1997).

the cell becomes, depends on its epigenetic circumstances. Moreover, the manner in which neurons from one area, such as the thalamus, connect to cells in the cortex depends very much on epigenetic circumstances, e.g., on the spontaneous activity, and later, the experience-driven activity, of the thalamic and cortical neurons. This is not to say that there are no causally significant differences between, for instance, the neonatal sucking reflex and knowing how to make a fire. Differences, obviously, there are. The essential point is that the differences do not sort themselves into the archaic 'nature' versus 'nurture' bins. Genes and extragenetic factors collaborate in a complex interdependency.[5]

Recent discoveries in neuropsychology point in this same direction. Hitherto, it was assumed that brain centers – modules dedicated to a specific task – were wired up at birth. The idea was that we were able to see because dedicated 'visual modules' in the cortex were wired for vision; we could feel because dedicated modules in the cortex were wired for touch, and so on.

The truth turns out to be much more puzzling.

For example, the visual cortex of a blind subject is recruited during the reading of braille, a distinctly nonvisual, tactile skill – whether the subject has acquired or congenital blindness. It turns out, moreover, that stimulating the subject's visual cortex with a magnet-induced current will temporarily impede his braille performance. Even more remarkably, activity in the visual cortex occurs even in normal seeing subjects who are blindfolded for a few days while learning to read braille.[6] So long as the blindfold remains firmly in place to prevent any light from falling on the retina, performance of braille reading steadily improves. The blindfold is essential, for normal visual stimuli that activate the visual cortex in the normal way impede acquisition of the tactile skill. For example, if after five days the blindfold is removed, even briefly while the subject watches a television program before going to sleep, his braille performance under blindfold the next day falls from its previous level. If the visual cortex can be recruited in the processing of nonvisual signals, what sense can we make of the notion of the dedicated vision module, and of the dedicated-modules hypothesis more generally?

What is clear is that the nature versus nurture dichotomy is more of a liability than an asset in framing the inquiry into the origin of plasticity in human brains. Its inadequacy is rather like the inadequacy of 'good versus evil' as a framework for understanding the complexity of political life in human societies. It is not that there is nothing to it. But it is like using a grub hoe to remove a splinter.

An appealing idea is that if you learn something, such as how to tie a trucker's knot, then that information will be stored in one particular location in the brain, along with related knowledge – say, between reef knots and half-hitches. That is, after all, a good method for storing tools and paper files – in a particular drawer at a particular location. But this is not the brain's way, as Karl Lashley first demonstrated in the 1920s.

5 See also Steven Quartz and Terrence J. Sejnowski, *Liars, Lovers, and Heroes* (New York: William Morrow, 2002).

6 See Alvaro Pascual-Leone et al., "Study and Modulation of Human Cortical Excitability with Transcranial Magnetic Stimulation," *Journal of Clinical Neurophysiology* 15 (1998): 333–343.

Lashley reasoned that if a rat learned something, such as a route through a certain maze, and if that information was stored in a single, punctate location, then you should be able to extract it by lesioning the rat's brain in the right place. Lashley trained twenty rats on his maze. Next he removed a different area of cortex from each animal, and allowed the rats time to recover. He then retested each one to see which lesion removed knowledge of the maze. Lashley discovered that a rat's knowledge could not be localized to any single region; it appeared that all of the rats were somewhat impaired and yet somewhat competent – although more extensive tissue removal produced more serious memory deficit.

As improved experimental protocols later showed, Lashley's non-localization conclusion was essentially correct. There is no such thing as a dedicated memory organ in the brain; information is not stored on the filing cabinet model at all, but distributed across neurons.

A general understanding of what it means for information to be distributed over neurons in a network has emerged from computer models. The basic idea is that artificial neurons in a network, by virtue of their connections to other artificial neurons and of the variable strengths of those connections, can produce a pattern that represents something – such as a male face or a female face, or the face of Churchill. The connection strengths vary as the artificial network goes through a training phase, during which it gets feedback about the adequacy of its representations given its input. But many details of how actual neural nets – as opposed to computer-simulated ones – store and distribute information have not yet been pinned down, and so computer models and neural experiments are coevolving.

Neuroscientists are trying to understand the structure of learning by using a variety of research strategies. One strategy consists of tracking down experience-dependent changes at the level of the neuron to find out what precisely changes, when, and why. Another strategy involves learning on a larger scale: what happens in behavior and in particular brain subsystems when there are lesions, or during development, or when the subject performs a memory task while in a scanner, or, in the case of experimental animals, when certain genes are knocked out? At this level of inquiry, psychology, neuroscience, and molecular biology closely interact.

Network-level research aims to straddle the gap between the systems and the neuronal levels. One challenge is to understand how distinct local changes in many different neurons yield a coherent global, system-level change and a task-suitable modification of behavior. How do diverse and far-flung changes in the brain underlie an improved golf swing or a better knowledge of quantum mechanics?

What kinds of experience-dependent modifications occur in the brain? From one day to the next, the neurons that collectively make me what I am undergo many structural changes: new branches can sprout, existing branches can extend, and new receptor sites for neurochemical signals can come into being. On the other hand, pruning could decrease branches, and therewith decrease the number of synaptic connections between neurons. Or the synapses on remaining branches could be shut down altogether. Or the whole cell might die, taking with it all the synapses it formerly supported. Or, finally, in certain special regions, a whole new neuron might be born and begin to establish synaptic connections in its region.

And that is not all. Repeated high rates of synaptic firing (spiking) will deplete the neurotransmitter vesicles available for release, thus constituting a kind of memory on the order of two to three seconds. The constituents of particular neurons, the number of vesicles released per spike, and the number of transmitter molecules contained in each vesicle, can change. And yet, somehow, my skills remain much the same, and my autobiographical memories remain intact, even though my brain is never exactly the same from day to day, or even from minute to minute.

No 'bandleader' neurons exist to ensure that diverse changes within neurons and across neuronal populations are properly orchestrated and collectively reflect the lessons of experience. Nevertheless, several general assumptions guide research. For convenience, the broad range of neuronal modifiability can be condensed by referring simply to the modification of synapses. The decision to modify synapses can be made either globally (broadcast widely) or locally (targeting specific synapses). If made globally, then the signal for change will be permissive, in effect saying, "You may change yourself now" – but not dictating exactly where or by how much or in what direction. If local, the decision will likely conform to a rule such as this: If distinct but simultaneous input signals cause the receiving neuron to respond with a spike, then strengthen the connection between the input neurons and the output neurons. On its own, a signal from one presynaptic (sending) neuron is unlikely to cause the postsynaptic (receiving) neuron to spike. But if two distinct presynaptic neurons – perhaps one from the auditory system and one from the somatosensory system – connect to the same postsynaptic neuron at the same time, then the receiving neuron is more likely to spike. This joint input activity creates a larger postsynaptic effect, triggering a cascade of events inside the neuron that strengthens the synapse. This general arrangement allows for distinct but associated world events (e.g., blue flower and plenty of nectar) to be modeled by associated neuronal events.

The nervous system enables animals to make predictions.[7] Unlike plants, animals can use past correlations between classes of events (e.g., between red cherries and a satisfying taste) to judge the probability of future correlations. A central part of learning thus involves computing which specific properties predict the presence of which desirable effects. We correlate variable rewards with a feature to some degree of probability, so good predictions will reflect both the expected value of the reward and the probability of the reward's occurring; this is the expected utility. Humans and bees alike, in the normal course of the business of life, compute expected utility, and some neuronal details are beginning to emerge to explain how our brains do this.

To the casual observer, bees seem to visit flowers for nectar on a willy-nilly basis. Closer observation, however, reveals that they forage methodically. Not only do bees tend to remember which individual flowers they have already visited, but in a field of mixed flowers with varying amounts of nectar they also learn to optimize their foraging strategy, so that they get the most nectar for the least effort.

Suppose you stock a small field with two sets of plastic flowers – yellow and blue – each with wells in the center into which precise amounts of sucrose have

7 John Morgan Allman, *Evolving Brains* (New York: Scientific American Library, 1999).

been deposited.[8] These flowers are randomly distributed around the enclosed field and then baited with measured volumes of 'nectar': all blue flowers have two milliliters; one-third of the yellow flowers have six milliliters, two-thirds have none. This sucrose distribution ensures that the mean value of visiting a population of blue flowers is the same as that of visiting the yellow flowers, though the yellow flowers are more uncertain than the blues.

After an initial random sampling of the flowers, the bees quickly fall into a pattern of going to the blue flowers 85 percent of the time. You can change their foraging pattern by raising the mean value of the yellow flowers – for example, by baiting one-third of them with ten milliliters. The behavior of the bees displays a kind of trade-off between the reliability of the source type and the nectar volume of the source type, with the bees showing a mild preference for reliability. What is interesting is this: depending on the reward profile taken in a sample of visits, the bees revise their strategy. The bees appear to be calculating expected utility. How do bees – *mere* bees – do this?

In the bee brain there is a neuron, though itself neither sensory nor motor, that responds positively to reward. This neuron, called VUMmx1 ('vum' for short), projects very diffusely in the bee brain, reaching both sensory and motor regions, as it mediates reinforcement learning. Using an artificial neural network, Read Montague and Peter Dayan discovered that the activity of vum represents prediction error – that is, the difference between 'the goodies expected' and 'the goodies received this time.'[9] Vum's output is the release of a neuromodulator that targets a variety of cells, including those responsible for action selection. If that neuromodulator also acts on the synapses connecting the sensory neurons to vum, then the synapses will get stronger, depending on whether the vum calculates 'worse than expected' (less neuromodulator) or 'better than expected' (more neuromodulator). Assuming that the Montague-Dayan model is correct, then a surprisingly simple circuit, operating according to a fairly simple weight-modification algorithm, underlies the bee's adaptability to foraging conditions.

Dependency relations between phenomena can be very complex. In much of life, dependencies are conditional and probabilistic: *If* I put a fresh worm on the hook, and *if* it is early afternoon, then *very probably* I will catch a trout *here*. As we learn more about the complexities of the world, we 'upgrade' our representations of dependency relations;[10] we learn, for example, that trout are more likely to be caught when the water is cool, that shadowy pools are more promising fish havens than sunny pools, and that talking to the worm, entreating the trout, or wearing a 'lucky' hat makes no difference. Part of what we call intelligence in humans and other animals is the capacity to acquire an increasingly complex understanding of dependency relations. This allows us to distinguish

8 This experiment was done by Leslie Real, "Animal Choice Behavior and the Evolution of Cognitive Architecture," *Science* (1991): 980–986.

9 See Read Montague and Peter Dayan, "Neurobiological Modeling," in William Bechtel, George Graham, and D. A. Balota, eds., *A Companion to Cognitive Science* (Malden, Mass.: Blackwell, 1998).

10 Clark N. Glymour, *The Mind's Arrows* (Cambridge, Mass.: MIT Press, 2001). See also Alison Gopnik, Andrew N. Meltzoff, and Patricia K. Kuhl, *The Scientist in the Crib* (New York: William Morrow & Co., 1999).

fortuitous correlations that are not genuinely predictive in the long run (e.g., breaking a tooth on Friday the thirteenth) from causal correlations that are (e.g., breaking a tooth and chewing hard candy). This means that we can replace superstitious hypotheses with those that pass empirical muster.

Like the bee, humans and other animals have a reward system that mediates learning about how the world works. There are neurons in the mammalian brain that, like yum, respond to reward.[11] They shift their responsiveness to a stimulus that predicts reward, or indicates error if the reward is not forthcoming. These neurons project from a brainstem structure (the ventral tegmental area, or 'VTA') to the frontal cortex, and release dopamine onto the postsynaptic neurons. The dopamine, only one of the neurochemicals involved in the reward system, modulates the excitability of the target neurons to the neurotransmitters, thus setting up the conditions for local learning of specific associations.

Reinforcing a behavior by increasing pleasure and decreasing anxiety and pain works very efficiently. Nevertheless, such a system can be hijacked by plant-derived molecules whose behavior mimics the brain's own reward system neurochemicals. Changes in reward system pathways occur after administration of cocaine, nicotine, or opiates, all of which bind to receptor sites on neurons and are similar to the brain's own peptides. The precise role in brain function of the large number of brain peptides is one of neuroscience's continuing conundrums.[12]

11 See Paul W. Glimcher, *Decisions, Uncertainty, and the Brain* (Cambridge, Mass.: MIT Press, 2003).

12 I am grateful to Roger Guillemain for discussing this point with me.

These discoveries open the door to understanding the neural organization underlying prediction. They begin to forge the explanatory bridge between experience-dependent changes in single neurons and experience-dependent guidance of behavior. And they have begun to expose the neurobiology of addiction. A complementary line of research, meanwhile, is untangling the mechanisms for predicting what is nasty. Although aversive learning depends upon a different set of structures and networks than does reinforcement learning, here too the critical modifications happen at the level of individual neurons, and these local modifications are coordinated across neuronal populations and integrated across time.

Within other areas of learning research, comparable explanatory threads are beginning to tie together the many levels of nervous system organization. This research has deepened our understanding of working memory (holding information at the ready during the absence of relevant stimuli) spatial learning, autobiographical memory, motor skills, and logical inference. Granting the extraordinary research accomplishments in the neuroscience of knowledge, nevertheless it is vital to realize that these are still very early days for neuroscience. Many surprises – and even a revolution or two – are undoubtedly in store.

Together, neuroscience, psychology, embryology, and molecular biology are teaching us about ourselves as *knowers* – about what it is to know, learn, remember, and forget. But not all philosophers embrace these developments as progress.[13] Some believe that what we call

13 I take it as a sign of the backwardness of academic philosophy that one of its most esteemed living practitioners, Jerry Fodor, is widely sup-

external reality is naught but an idea created in a nonphysical mind, a mind that can be understood only through introspection and reflection. To these philosophers, developments in cognitive neuroscience seem, at best, irrelevant.

The element of truth in these philosophers' approach is their hunch that the mind is not just a passive canvas on which reality paints. Indeed, we know that brains are continually organizing, structuring, extracting, and creating. As a central part of their predictive functions, nervous systems are rigged to make a coherent story of whatever input they get. 'Coherencing,' as I call it, sometimes entails seeing a fragment as a whole, or a contour where none exists; sometimes it involves predicting the imminent perception of an object as yet unperceived. As a result of learning, brains come to recognize a stimulus as indicating the onset of meningitis in a child, or an eclipse of the Sun by the Earth's shadow. Such knowledge depends upon stacks upon stacks of neural networks. There is no apprehending the nature of reality except via brains, and via the theories and artifacts that brains devise and interpret.

From this it does not follow, however, that reality is *only* a mind-created idea. It means, rather, that our brains have to keep plugging along, trying to devise hypotheses that more accurately map the causal structure of reality. We build the next generation of theories upon the scaffolding – or the ruins – of the last. How do we know whether our hypotheses are increasingly adequate? Only by their relative success in predicting and explaining.

But does all of this mean that there is a kind of fatal circularity in neuroscience – that the brain necessarily uses itself to study itself? Not if you think about it. The brain I study is seldom my own, but that of other animals or humans, and I can reliably generalize to my own case. Neuroepistemology involves many brains – correcting each other, testing each other, and building models that can be rated as better or worse in characterizing the neural world.

Is there anything left for the philosopher to do? For the neurophilosopher, at least, questions abound: about the integration of distinct memory systems, the nature of representation, the nature of reasoning and rationality, how information is used to make decisions, what nervous systems interpret as information, and so on. These are questions with deep roots reaching back to the ancient Greeks, with ramifying branches extending throughout the history and philosophy of Western thought. They are questions where experiment and theoretical insight must jointly conspire, where creativity in experimental design and creativity in theoretical speculation must egg each other on to unforeseen discoveries.[14]

14 Many thanks to Ed McAmis and Paul Churchland for their ideas and revisions.

---

ported for the following conviction: "If you want to know about the mind, study the mind – not the brain, and certainly not the genes" (*Times Literary Supplement,* 16 May 2003, 1 – 2). If philosophy is to have a future, it will have to do better than that.