3 The nature of learning and the functional architecture of the brain

Charles R. Gallistel

Assumptions about the nature of learning and assumptions about the functional architecture of the brain are intimately intertwined in contemporary cognitive science and neuroscience. There are two radically different analytic frameworks for thinking about learning; for each, there is a corresponding functional architecture imputed to the brain.

On the first analysis, learning is the extraction from experience of information about the world, which is carried forward in memory to inform future behavior. On the second analysis, learning is the remolding of a plastic brain by experience so as to make the brain better adapted to the experienced world.

On the first analysis, the brain has the functional architecture of a Turing machine (Turing, 1936, see Figure 3.1A), that is, of an information-processing device, such as a conventional computer. The fundamental components of this architecture are a read/write memory (represented by the tape in Figure 3.1A) and compositional operators (represented by the addition register in Figure 3.1A).

The memory contains information about the world in the form of symbols. In a conventional computer, these are bit patterns, strings of 1s and 0s. They represent facts about the world, such as the width of a brick wall. Symbols representing facts about the world are the data on which a data-processing device like a computer operates. The essential function of the memory is the carrying of information forward in time. It is the repository where information resides when it is not in use.

Compositional operators compose new data structures or expand old data structures, by operating on data read from memory. Data structures are sets of related symbols, stored in memory in such a way as to encode the relations that obtain between them. An example would be a pair of symbols, representing the width and height of a wall. The pair is stored in such a way as to indicate which symbol represents the width, which the height, and which wall it is whose width and height they represent.

Composition means putting together. It is the essence of data processing. Multiplying two numerical symbols – for example, the symbols representing the width and height of the wall to obtain a symbol representing the area of the wall – is an example of a compositional operation. It puts together two

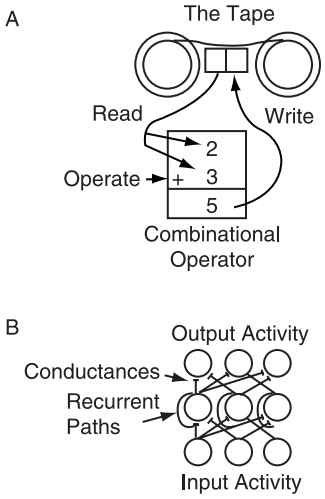

Figure 3.1 A. The functional architecture of an information-processing system. The Tape is the read/write memory element with which Turing endowed his abstraction of the essential elements of a universal computing machine. The arithmetic summation is an example of the symbol-combining (compositional) operations that constitute the other essential element. B. The functional architecture of recurrent switching net.

input symbols to generate a third symbol, their product. When this symbol is written into the right place in memory – taking its place among the other wall-related symbols – the data structure representing the wall is thereby expanded.

There is no memory in the conventional computer-science sense on the second analysis, the analysis inspired by current ideas about the structure and function of the nervous system. On this second analysis, the brain has the functional architecture of a recurrent switching network (Rumelhart & McClelland, 1986, see Figure 3.1B). The fundamental components of this architecture are nodes (pseudo-neurons) with variable activity levels. The nodes are connected to each other by graded and signed conductances, which determine the flow of activity between nodes. Critical components of this internodal connectivity are recurrent pathways, which conduct activity from a node back to itself. Without them there is no sustained activity in the network because, as Sherrington said of the nervous system: "[the function of a switching network] can be summed up in one word, *conduction*" (Sherrington, 1947).

What comes naturally

The two functional architectures differ strikingly in what they make it natural (easy, straightforward) for a brain to do. On the first analysis, what comes

naturally is the composing of data structures: copying them, concatenating them, multiplying, dividing, adding, subtracting them, and so on, to create new data structures and new values for old ones. On this analysis, the brain represents the world by composing data structures that refer to selected aspects of it.

On the second analysis, what comes naturally is classification. A recurrent switching network settles into different stable patterns of activity (attractor states) depending on which input it gets. The number of distinct stable attractor states is, however, much less than the number of distinct inputs. Thus, many different inputs cause the system to settle into any given stable attractor state. This seems analogous to a category and its instances: instances (inputs) differ but they activate the same category (attractor state). The relation between inputs and the attractor states they activate is determined by the conductances in the switching network. The conductances are functions of a learning rule of some kind applied to the animal's experience. The conductance between a signal-sending node and a signal-receiving node determines the magnitude (and sign) of the effect that a signal from the sending node has on receiving node. Thus, what it is easy or natural for a recurrent switching network to do is to learn to categorize states of the world. A recurrent network represents the world by categorizing it.

The two frameworks also differ in their assumptions about how the brain's activity becomes manifest in behavior. On the first analysis, behavior-generating programs get control and decision variables from the information processing system. They compare them to action-specific decision criteria (thresholds) to decide on actions, which are then parameterized by the control variables. The distinction between the information-processing component, which extracts information about the world from input, and the action-generating system, which makes use of the information, is generally readily apparent in this kind of architecture. On the second analysis, the distinction between these two roles is less clear. Often, the activity pattern on the output side of the network is taken, at least implicitly, to specify an action appropriate to the specified category.

Courts of last appeal

Partisans of the different frameworks differ in their courts of last appeal. For partisans of the first analysis, the court of last appeal is behavior: they argue, as I will argue, that the behavioral facts imply an information-processing architecture. There is no way to explain the behavioral facts except by assuming a read/write memory and the composition of data structures, both of which functionalities are absent in the second kind of architecture. Therefore, we must assume the existence of mechanisms in the nervous system that perform these functions, despite the fact that what we currently know about the nervous system does not support such an assumption. On this analysis, what we know about behavior is a guide to what we must look for in the nervous system.

For partisans of the second analysis, the court of last appeal is the nervous system. It manifestly has, they argue, the architecture they assume in modeling behavior. Open any textbook on the nervous system and what you find in the description of the system's observable structure and in the records of its electrical activity is the functional structure portrayed in Figure 3.1B. Nowhere do you find a description of a mechanism with the properties of a read/write memory. You find, it is true, descriptions of localized activity in anatomical spaces that map real-world spaces. For example, the location of activity in the deep layers of the superior colliculus (a two-dimensional anatomical space) maps the location of distal saccade targets in a two-dimensional directional space centered on the optic axis of the eye (Sparks & Groh, 1995). These localized activities look like candidate data structures; they look like vectors in the nervous system that refer to and may therefore represent vector variables in the world outside the brain. If, however, that is the appropriate interpretation of them, then something crucial is missing, namely a description of the mechanisms or processes that compose these data structures to create new data structures. How are they added, subtracted, multiplied, copied, concatenated, subordinated, and so on? If the nervous system is a system for composing data structures, what are its built-in compositional operations? On this question, the texts are also silent. Partisans of the second analysis take these silences as evidence of absence: what has not been described presumably does not exist and should not be assumed in making models meant to explain behavior.

The behavioral case

Many simple and robust behavioral phenomena seem inexplicable – or very awkwardly explicable – in the absence of a read/write memory and the composition of data structures. I briefly review three of them.

Dead reckoning

Experiment has shown that animals keep track of their position and heading relative to their nest and other points of interest in the world by integrating their velocity with respect to time, or, equivalently, by summing successive small displacement vectors to get the net displacement vector (Alyan & Jander, 1994; Alyan & McNaughton, 1999; Durier & Rivault, 1999; Etienne et al., 1991; Georgakopoulos & Etienne, 1994; Mittelstaedt & Mittelstaedt, 1980; Save et al., 1998; Schatz et al., 1999; Séguinot et al., 1993; Wehner & Wehner, 1986; Wohlgemuth et al., 2001).

Dead reckoning is dead easy in a device with the architecture shown in Figure 3.1A. A displacement vector is a simple data structure consisting of two symbols that together specify a small change in position, for example <0.9, 0.4>. The first of the two symbols represents the distance the animal has just moved in one of two perpendicular directions (e.g., north); the other

represents the distance it has just moved in the other direction (e.g., east). Each small movement (change in position) generates a displacement vector. The running sum of these vectors – the sum of all the symbols in the first position together with the sum of all the symbols in the second position – is the net displacement, which represents where the animal is relative to where it started. (This running sum is often called the home vector.)

To implement dead reckoning in a system with the architecture of Figure 3.1B is awkward (Samsonovich & McNaughton, 1997). As Samuel Johnson said of a dog walking on its hind legs, "It is not done well, but you are surprised that it is done at all."[1]

The dance of the foraging bee

The foraging bee returning from a rich food source does a dance that specifies the direction and distance of the source (Frisch, 1967). The dance is in the form of a figure 8, with the 'waggle' run constituting the common portion of the two loops. The angle of the waggle run with respect to vertical specifies the solar bearing of the source (its direction relative to the sun) and the number of waggles specifies the distance.

There are no implementational mysteries here, if we grant the bee's nervous system the functional architecture in Figure 3.1A. When the bee is at the source, its dead-reckoning vector specifies the source's location relative to the hive (a vector structure). As the bee ingests the nectar preparatory to carrying it back to the hive, it gets a sensory input specifying its richness (a scalar structure). It writes the location vector and the richness scalar to memory. When it reaches the hive, the dance program retrieves these data structures (reads the memory). It uses the scalar that represents the richness as a decision variable; only if its value exceeds a threshold (decision criterion) does the bee dance. If it does dance, the remembered location vector at the source parameterizes the behavior. Its angle specifies the direction of the waggle run, while its magnitude specifies the number of waggles.

How to implement this within the architecture in Figure 3.1B is another matter. Suppose we grant, as assumed in Samsonovich and McNaughton (1997), that several interconnected nets with purpose-specific structures give rise to activity patterns whose locations within the nets specify the bee's location when it is at the source. Now what? The net has no read/write memory, no way of preserving for future use the information about the world expressed in its current activity. If we shut off input to these nets to maintain the current activity patterns for future use, how does the bee find its way home? The activity patterns in these nets represent where the bee is at the moment in relation to the hive. That information is presumably crucial for the moment-to-moment control of the homeward flight, although how this control is effected is not specified in the model.

What neural nets have trouble accommodating is the fact that the information that informs behavior is not delivered to the brain all at once, nor at the

time it is needed to inform behavior; it arrives piecemeal over time and often long before it is needed. This elementary truth about the character of experience is why the brain needs a read/write memory. Without it, there is no way to make the information gained from current experience accessible to future computation and behavioral control. Put another way, there is no way to compose the activity patterns in the output nodes of the architecture in Figure 3.1B, and that is a devastating shortcoming.

Temporal learning in conditioning experiments

In training phases of Pavlovian conditioning experiments, stimuli called conditioned stimuli (CSs for short), which are initially motivationally neutral (e.g., noises and lights), and unconditioned stimuli (USs for short), which are behaviorally motivating (e.g., foot shock), are repeatedly presented in one or more temporal relations to one another. After the training phases, the subject's reaction to the previously neutral stimuli is measured on test trials.

The temporal relations among the CSs and the US during the training phases determine the motivational valence of the subject's response to the CSs on the test trials. If, for example, a light regularly followed foot shock during one phase of training, the subject's reaction to the light on test trials involves the inhibition of the fear aroused by the anticipation of shock. This is usually explained by assuming that the conditioning experience has created an inhibitory connection (negatively signed conductance) between the node(s) activated by the light and the nodes whose activity constitutes the fear state, although what the learning rule is that creates a forward conductance (from the light node to the fear node) in response to the backward temporal pairing of the node-activating inputs (shock first, then noise) is not clear.

If in earlier phase of the training, the light was regularly preceded by a noise, and if the interval from noise to light in this phase was longer than the interval from shock to light in the latter phase of training, then when tested with the noise, the subject reacts fearfully to it (Barnet et al., 1997; Barnet & Miller, 1996 see Figure 3.2). This is harder to understand from the perspective of conductance changes within a switching network. The activity conducted from the node activated by light to the node whose activity constitutes fear is negatively signed. Activity in the light node reduces activity in the fear node. Because the excitation of the noise node regularly preceded the activation of the light node during the phase of training when they were paired, the sign of the conductance between them is presumably positive, which means that activity in the noise node increases activity in the light node. If activity in the noise node increases activity in the light node and activity in the light node decreases activity in the fear node, why does activating the noise node excite the fear node? It ought to inhibit it.

This latter result is much easier to understand if the brain has a compositional architecture with a read/write memory. From this perspective, the

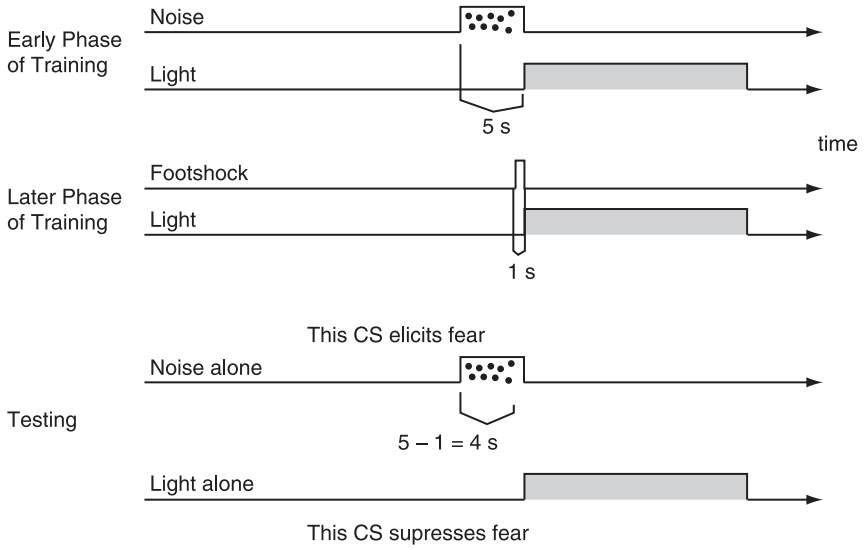

Figure 3.2 Simplified schematic of a Pavlovian conditioning experiment by Barnet and Miller (1996). During the first phase of training, the light (tint) followed the noise (dots) by 5 s. During the second phase of training, shock preceded the light by 1 s. When the two CSs (noise and light) were tested individually, following the second phase of training the light suppressed fear. The noise, which was paired only with the light during training, never with the shock, nonetheless elicited fear. The expected interval between the onset of noise and the onset of shock is 4 s, on the assumption that the brain can compose the remembered values of the noise–light interval (5 s) and the shock–light interval (–1 s).

training phase with the noise followed by the light created a data structure with three fields: one contained a vector specifying the properties of the noise, one a vector specifying the properties of the light, and one a scalar (single quantity) specifying the temporal interval between their onsets. We may represent this data structure as follows: {<noise>, interval, <light>}. The training phase with the shock followed by the noise created a second such data structure: {<shock>, interval, <light>}. The brain composed the two structures to generate a third such structure: {<noise>, interval, <shock>}. The scalar in the middle field of this third structure is the first interval minus the second interval. It represents the expected interval to shock onset following noise onset. This computed temporal expectation explains the subject's fearful reaction to the noise.

Here again, the compositional character of brain activity is evident in the behavior it generates. All of these examples are simple, and they are but a small sampling of the many well-established results from behavioral experiments that are unintelligible unless one assumes a read/write memory and a set of compositional operators (Gallistel, 1990; Gallistel & Gibbon, 2000,

2002). If behavior is the last court of appeal, then there are mechanisms in the nervous system not yet dreamed of in the philosophy of neuroscientists.

This is not the first time this state of affairs has arisen in the history of transdiscipline scientific reductionism. Lord Kelvin famously explained to Darwin that the enormous age for the earth implied by his theory of evolution was not consistent with what physicists knew about the conceivable duration of heat-generation processes in a body the size of the sun (Burchfield, 1990). Kelvin did not know what he did not know. In this case, what he did not know was radioactivity. Its discovery came many years later.

Before Watson and Crick's (1953) epoch-making deduction of the molecular structure of the gene, many biochemists doubted that there were genes (Judson, 1980). The geneticists' concept of a gene was biochemically unintelligible. The gene had two properties for which no plausible biochemical mechanism could be suggested. It was a (putatively) molecular structure that could somehow make a copy of itself. Second, it could somehow determine the sequence of the amino acids in the synthesis of an enzyme. Again, the doubters did not know what they did not know. When the revelation came, it created a new discipline, molecular biology, built on concepts foreign to the biochemistry that preceded this discovery.

These two examples urge caution on those who would disregard the seemingly clear neuroscientific implications of the behavioral data.

Note

1 Boswell, J. (1763). *Life of Johnson 1 Vol. ii. Chap. ix*.

References

Alyan, S., & Jander, R. (1994). Short-range homing in the house mouse, *Mus musculus:* Stages in the learning of directions. *Animal Behaviour*, *48*, 285–298.

Alyan, S., & McNaughton, B.L. (1999). Hippocampectomized rats are capable of homing by path integration. *Behavioral Neuroscience*, *113*(1), 19–31.

Barnet, R.C., Cole, R.P. et al. (1997). Temporal integration in second-order conditioning and sensory preconditioning. *Animal Learning and Behavior*, *25*(2), 221–233.

Barnet, R.C., & Miller, R.R. (1996). Second order excitation mediated by a backward conditioned inhibitor. *Journal of Experimental Psychology: Animal Behavior Processes*, *22*(3), 279–296.

Burchfield, J.D. (1990). *Lord Kelvin and the age of the Earth*. Chicago: University of Chicago Press.

Durier, V., & Rivault, C. (1999). Path integration in cockroach larvae, *Blattella germanica* (L.) (insect: Dictyoptera): Direciton and distance estimation. *Animal Learning and Behavior*, *27*(1), 108–118.

Etienne, A.S., Hurni, C. et al. (1991). Twofold path integration during hoarding in the golden hamster. *Ethology, Ecology, Evolution*, *3*, 1–11.

Frisch, K.V. (1967). *The dance-language and orientation of bees*. Cambridge, MA: Harvard University Press.

Gallistel, C.R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/ MIT Press.

Gallistel, C.R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review*, *107*, 289–344.

Gallistel, C.R., & Gibbon, J. (2002). *The symbolic foundations of conditioned behavior*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Georgakopoulos, J., & Etienne, A. (1994). Identifying location by dead reckoning and extenal cue. *Behavioural Processes*, *31*, 57–74.

Judson, H. (1980). *The eighth day of creation*. New York: Simon & Schuster.

Mittelstaedt, M.L., & Mittelstaedt, H. (1980). Homing by path integration in a mammal. *Naturwissenschaft*, *67*, 566–567.

Rumelhart, D.E., & McClelland, J.L. (Eds.). (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Samsonovich, A., & McNaughton, B.L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, *17*, 5900–5920.

Save, E., Cressant, A. et al. (1998). Spatial firing of hippocampal place cells in blind rats. *Journal of Neuroscience*, *18*(5), 1818–1826.

Schatz, B., Chameron, S. et al. (1999). The use of path integration to guide route learning in ants. *Nature*, *399*, 769–772.

Séguinot, V., Maurer, R. et al. (1993). Dead reckoning in a small mammal: The evaluation of distance. *Journal of Comparative Physiology, A*, *173*, 103–113.

Sherrington, C.S. (1947). *The integrative action of the nervous system*. New Haven, CT: Yale University Press. (First published 1906.)

Sparks, D.L., & Groh, J.F. (1995). The superior colliculus: A window for viewing issues in integrative neuroscience. In M.S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 565–584). Cambridge, MA: MIT Press.

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society 2nd ser.*, *42*, 230–265.

Watson, J.D., & Crick, F.H.C. (1953). Molecular structure of nucleic acids. *Nature*, *171*, 737–738.

Wehner, R., & Wehner, S. (1986). Path integration in desert ants: Approaching a long-standing puzzle in insect navigation. *Monitore Zoologiea Italiana*, *20*, 309–331.

Wohlgemuth, S., Ronacher, B. et al. (2001). Ant odometry in the third dimension. *Nature*, *411*, 795–798.