Chapter 5

Connectionist Models of Human Associative Learning

A. J. Wills

Warning: If you are new to the study of human associative learning and have skipped the first four chapters you're going to find the next four pretty tough. This is because all these chapters, particularly this one, draw on concepts introduced in chapters 1–4.

In chapter 1, I drew a distinction between two quite different concepts that sometimes both attract the term *associative learning*. On some occasions, "associative learning" is used to define a particular type of problem that an organism has to solve. On other occasions, "associative learning" is used as a theoretical statement about the sorts of mental processes by which the organism solves this type of problem.

ASSOCIATIVE LEARNING → LEARNING OF ASSOCIATIONS

Let's consider "associative learning" as a problem definition first. In this sense of the term, "learning" refers to a relatively permanent change in response potentiality caused by information available from the organism's perceptual receptors. This rather technical definition is a variant of the definition offered by Reber (1985, p. 395).

Breaking this definition of "learning" down into its constituent components, "response potentiality" indicates that learning results in a potential to respond differently. Sometimes the organism actually will act differently. At other times, no immediate behavioral change is observed but evidence that learning has occurred emerges later. The use of "response potentiality" also underlines the important idea that learning is a hypothetical event for which behavior provides evidence.

The phrase "relatively permanent change" is intended to exclude various types of momentary changes in response potentiality. This is a fuzzy boundary, but changes in response potentiality that persist for no more than a few hundred milliseconds are not generally considered as

"learning." The phrase "caused by information available from the organism's perceptual receptors" is a statement about the data upon which learning operates. "Perceptual receptors" indicates structures such as the retina, the cochlea, the somatosensory receptors, and so on. One important aspect of this definition is that it is not intended as a statement of process. Hence, no assumption is being made about an upper limit to the amount of processing that results from sensory information, the time course over which that happens, or the extent to which the learning process importantly involves integration with already learned information.

When "associative learning" is used as a definition of a type of problem facing an organism, the intent behind this phrase is perhaps better expressed as "the learning of one or more associations." The intent behind the term *association* is statistical; it is the extent to which changes in one environmental variable are related to changes in another. The provision of an appropriate statistical measure of association is not a trivial problem. For example, one is likely to use a different measure depending on whether a predictive or correlational relation is being considered. A predictive relation has a particular direction. For example, if you know a car has a dead battery you can predict pretty reliably that the car will not start. However, if a car will not start, you should be much less confident about predicting the presence of a dead battery. Delta P (chaps. 2 & 4) is one example of a measure of a predictive associative relationship. In contrast, a correlational relationship as measured by, for example, Pearson's r is bidirectional. Even within these two classes (predictive and correlational) the choice of statistic is not straightforward. Recall, for example, from chapter 2 that the delta P and PowerPC equations provide two different potential measures of the strength of a predictive relationship.

ASSOCIATIVE LEARNING → CONNECTIONISM

Now let's turn to the usage of "associative learning" in the sense of a class of theory about the processes involved in the "learning of associations." Here we can haul ourselves out of the linguistic treacle by using the term *connectionism*. As I said in chapter 1, there is an unfortunate tendency when discussing the history of psychology to assume that connectionism started in 1986 with the publication of the PDP manuals (Rumelhart, McClelland, & The PDP Research Group, 1986). These manuals undoubtedly had an enormous impact; they encouraged the mainstream of human cognitive psychology to reconsider the usefulness of connectionism as a theoretical system. However, connectionism is much more than 20 years old. The word *connectionism* can be traced back at least as far as Thorndike (1898), whereas the development of associationism can be traced from Aristotle, through the British Empiricists (e.g., Hume, 1739/1978), to Ebbinghaus (1885/1913) and Pavlov (1909/1928) and, from there, throughout the 20th century.

Connectionism is a theoretical approach that assumes learning results from the formation of connections between representations. These representations (often called "nodes") have a variable level of activation. The activation of a node is passed through all of its outward connections to other nodes, hence determining the activity level of those nodes. The connections between nodes have a variable "strength" or "weight." The stronger the connection, the more efficient it is at transmitting activation. The strength of connections is changed by a learning algorithm. Many different learning algorithms have been proposed, the simplest of which is probably the Hebbian algorithm (Hebb, 1949). In Hebbian learning, the connection weight between two representations increases if they are cojointly active.

For many, one of the appeals of a connectionist approach is that it appears to be a simplified model of the action of neurons. This perhaps gives the potential, in the long term, for more unified accounts of the human mind that incorporate both physiological and psychological observations. On the other hand, one of the most commonly used learning algorithms (Rumelhart, Hinton, & Williams, 1986) allows information to travel in both directions down the same connections; something that given our current understanding of neurophysiology seems rather unlikely. So, whereas some theorists do see the integration of physiology and psychology as an important goal, to others the neuron is more like a descriptive metaphor for the operations of a connectionist system. The implication of this must be that connectionist theories have (at least perceived) virtues other than the potential to integrate physiology and psychology.

Probably the other main appeal of connectionist models is the level of specificity that is gained. A theory expressed in connectionist terms seems to leave much less room for ambiguity and interpretation than a theory expressed in more informal terms. One example of an informally expressed theory is Alan Baddeley's visuo-spatial sketchpad (see, e.g., Baddeley, 1986). Yet, a theory can clearly reach a high level of mathematical specificity without being connectionist. For example, Ashby's accounts of categorical decisions (e.g., Ashby, 2000) are expressed specifically and mathematically but are not connectionist. Connectionist accounts can even have certain pragmatic disadvantages compared to some other kinds of mathematical model. For example, deriving predictions from connectionist systems can become quite involved because they are generally nonlinear systems (which makes the mathematics more complex). For similar reasons, it can also be difficult to determine whether a particular behavior of a connectionist system is a general property of the model or whether it is parameter-specific. Such complexity may be necessary to explain human behavior but it is not, in itself, a virtue.

Probably the best way to consider connectionist systems are as theoretical accounts that have a high level of specificity while simultaneously taking into account certain basic principles of neural

function. It is this combination that presumably leads to their continued popularity.

OUTLINE OF THIS SECTION

The following three chapters are examples of how connectionist models can be employed as theories of human associative learning. All three chapters are broadly similar in approach, probably because four of the five authors have worked closely with each other for some years (Mark Suret, Mike Le Pelley, and I were all members of Ian McLaren's research group for a number of years during the period 1994–2003). As a result of these close links, the level of agreement across these chapters is greater than it is across the field of connectionist modeling of human associative learning as a whole.

Chapter 6 starts with an introduction to connectionism and describes the Hebbian learning algorithm alluded to in the previous section. Jan and I then continue with a consideration of how connectionist systems account for the learning that undoubtedly can occur in the absence of feedback. We then consider the Rescorla–Wagner rule (Rescorla & Wagner, 1972) but, in contrast to previous chapters, the concentration is on the general strengths and limitations of the rule as an algorithm for learning from feedback, rather than on its ability to predict specific experimental results. Next, Jan and I make a case for the need to combine feedback and no-feedback learning systems. A simple integrated model is proposed as one example of how this could be done, and is tested against the results of a novel experiment. The chapter closes with a consideration of some of the more obvious limitations of the integrated model proposed.

Chapter 7 reprises the Rescorla–Wagner model, this time drawing attention to its prediction that all cues present on a given trial are subject to the same change in associative strength (assuming the cues are of equal salience). Following a very elegant demonstration by Rescorla (2000) that this is not the case for rats, Mike and Ian demonstrate that it is also not true for humans performing an allergy prediction task. They then introduce their APECS model—a different type of connectionist system—and show that it can predict the results found. Next, they demonstrate the presence, in humans, of a related effect with absent-but-expected, rather than present, cues (retrospective revaluation, see chap. 3). The APECS model is also able to predict these entirely novel results. Finally they present a study (based on work by Lochmann & Wills, 2003) that indicates that the APECS model needs to be modified to explain certain predictive history effects. Predictive history is the idea that cues that have a history of being predictive form associations more quickly than those that have a history of being nonpredictive, even when the outcomes being learned about are novel.

In order to explain the results found, Mike and Ian suggest the adoption of the associability change processes proposed by Nick Mackintosh

(1975). The basic idea is that, in addition to their variable activity, nodes representing stimuli have a variable "associability" that modulates the rate at which associative links from this node change in strength. In the Mackintosh system, if a stimulus is a good predictor its associability increases whereas if it is a poor predictor its associability decreases.

In chapter 8, Mark and Ian continue the theme of associability processes. They start by introducing Lawrence's (1952) "Transfer along a continuum" (TAC) finding with rats. TAC, roughly stated, is a demonstration that training on an easy discrimination (e.g., black vs. white) before transferring to a difficult discrimination (e.g., light gray vs. dark gray) can result in better performance on the "hard" discrimination than an equivalent amount of training on the "hard" discrimination from the outset. The basic result can be explained without recourse to the concept of associability. However, a more sophisticated version of the experiment, again performed with rats (Mackintosh & Little, 1970), seems to require some kind of associability process. Mark and Ian demonstrate that effects analogous to those found by Mackintosh and Little can also be found in humans. They then go on to demonstrate how the McLaren and Mackintosh (2000, 2002) connectionist model can be modified to include the sort of associability process originally proposed by Mackintosh some 25 years earlier. They also demonstrate that this modified model can reproduce in detail the patterns of results found in their human experiments.

REFERENCES

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology, 44*, 310–329.

Baddeley, A. (1986). *Working memory.* New York: Oxford University Press.

Ebbinghaus, H. (1913). *Über das Gedächtnis* (H. Ruyer & C. E. Bussenius, Trans.). New York: Teachers College, Columbia University. (Original work published 1885)

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Hume, D. (1978). *A treatise of human nature.* New York: Oxford University Press. (Original work published 1739)

Lawrence, D. H. (1952). The transfer of a discrimination along a continuum. *Journal of Comparative and Physiological Psychology, 45*, 511–516.

Lochmann, T., & Wills, A. J. (2003). Predictive history in an allergy prediction task. In F. Schmalhofer, R. M. Young, & G. Katz (Eds.), *Proceedings of EuroCogSci03: The European Cognitive Science Conference* (pp. 217–222). Mahwah, NJ: Lawrence Erlbaum Associates.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298.

Mackintosh, N. J., & Little, L. (1970). An analysis of transfer along a continuum. *Canadian Journal of Psychology, 24*, 362–369.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior, 28*(3), 211–246.

McLaren, I. P. L., & Mackintosh, N. J. (2002). An elemental model of associative learning: II. Generalization and discrimination. *Animal Learning & Behavior, 30,* 177–200.

Pavlov, I. P. (1928). Natural science and the brain. In W. H. Gantt (Ed.), *Lectures on conditioned reflexes* (Vol. 1, pp. 120–130). London: Lawrence & Wishart. (Original work published 1909)

Reber, A. S. (1985). *Dictionary of psychology.* London: Penguin.

Rescorla, R. A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behavior Processes, 26,* 428–438.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research* (pp. 64–99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

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

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review, 8.*