1 Introduction

It is hard to think of any significant aspect of our lives that is not influenced by what we have learned in the past. The world looks and sounds the way it does because as infants we learned to partition it up in certain meaningful ways: we see familiar faces rather than meaningless blobs of colour and hear words rather than noise. Similarly, we behave in the ways we do because we have learned from past experience that our various actions have certain specific consequences. Like many topics of psychological inquiry, the importance of learning can perhaps best be realised by considering what life is like for people who have learning difficulties. Consider the case of Greg, a patient described by Sacks (1992), who became profoundly amnesic as a result of a benign brain tumour that was removed in 1976. Although his memory for events from his early life was almost completely normal, Greg remembered virtually nothing that had happened to him from 1970 onwards and appeared quite unable to learn anything new. He continued to believe, for instance, that Lyndon Johnson was the American President. In 1991 he was taken to a rock concert given by a group that he had been a great fan of in the 1960s, and despite sitting through the concert in rapture and recalling many of the songs, by the next morning he had no memory of the concert. More distressingly, when told of his father's death he was immeasurably sad but forgot the news within a few minutes. He was unable to learn that his father was no longer alive, and relived his grief anew every time he was told the news.

Another of Greg's difficulties, more mundane but also more representative of the sorts of behaviours studied by psychologists, was his inability to remember lists of words for more than a few minutes. This difficulty had nothing to do with understanding the words, since Greg's linguistic knowledge, having been learned early in life, was preserved. Rather, it is attributable to an inability to learn new associations between a certain context – the time and place in which the list was read – and the words on the list. For a normal person, remembering a list of words heard in a certain context would be relatively straightforward, since recollection of the context would bring the words to mind simply by association. Greg's other problems, such as his inability to learn who is President, can also be interpreted in terms of a basic difficulty in learning associations: a normal person will have little difficulty learning a new association between a name and the label 'President'.

Although learning lists of unrelated words may not be of much value in

the real world, in general this sort of *associative* learning is at the heart of any organism's psychological capabilities, because it endows the organism with the ability to adapt its behaviour as a result of acquiring information about associations or contingencies that exist between events in its environment. The ability to search out rewards like food and avoid threats like predators can only be achieved by learning predictive relations between rewards and threats, on the one hand, and events that are reliable signals of them, on the other. And as many researchers have observed, this adaptive ability is a major feature of what we understand by terms like 'intelligence' and 'intelligent behaviour.' Indeed, as Howard (1993) has pointed out, one widespread definition of intelligence equates it with flexible learning: a person (or animal) who is able to learn efficiently and transfer its knowledge to new situations that it encounters is called intelligent.

This book reviews research conducted over the last 20 or so years on the psychology of human learning and focuses specifically on associative learning. In an associative learning situation, the environment (or the experimenter) arranges a contingent relationship between events, allowing the person to predict one from the presence of others. The predictive events will either be external signals which I shall term 'cues', or the subject's own actions. Predictive relationships can be of two sorts, causal or structural. The most obvious form of relationship is *causal*, where one event or set of events is followed after an interval of time by another. For instance, in my office there is (barring an electrical fault) a consistent causal relationship between pressing the light switch and the light coming on. In contrast, we may say that a relationship is *structural* when an organism learns to predict one feature or attribute of an object or event from the presence of other features that regularly co-occur with it. For example, as a result of exposure to the co-occurrence of the sight and sound of running water, an organism may benefit from being able to predict that the sound of water is a good index of the sight of it. The ability to classify objects is another example of structural prediction. When I classify a particular sound as a word I am assigning it to a category, the category consisting of all the possible tokens of that spoken word. But the relationship between sound and category is structural, not causal: being a member of the category is a feature or property of the sound.

The term 'associative learning' has traditionally been meant to provide a contrast with 'nonassociative learning', but in fact this contrast is probably of little significance. Typically, the term nonassociative learning has been used to describe phenomena such as habituation, priming, and perceptual learning where, in contrast to associative situations, no explicit contingencies between experimenter-defined stimuli or actions are programmed, but where learning can nevertheless be observed. Thus in perceptual learning, subjects are simply exposed to isolated stimuli (such as faces) and learn to discriminate them better than would otherwise be possible, without there

being any overt contingency between these stimuli and other events. The problem with this definition is that there inevitably are structural contingencies amongst the *elements* of a stimulus, and so as researchers like McClelland and Rumelhart (1985) and McLaren, Kaye and Mackintosh (1989) have noted, so-called nonassociative learning may be grounded in associative learning of those contingencies. I shall not discuss in any detail tasks such as priming that are typically regarded as examples of nonassociative learning, but it is worth bearing in mind that the principles of associative learning may be perfectly applicable to nonassociative learning as well.

What exactly is meant by 'learning'? The definition of this apparently innocuous term has been a topic of passionate debate by psychologists. In their enthusiasm to rid the subject of mentalistic concepts, the behaviourists argued that learning must be observable and that therefore it should be equated with the emergence of new patterns of responding. When we say that a dog in a laboratory Pavlovian conditioning experiment has learned something about the relationship between a bell and food, what we mean is just that a new behaviour has been conditioned: the dog salivates to the bell, whereas previously it did not. On such a view, we should only use the term 'learning' if there is some observable change in behaviour, in which case the new behaviour *is* the learning.

However, there are at least two obvious problems with this definition. The first is that learning may occur without any concomitant change in behaviour: if a cue and an outcome such as shock are presented to subjects administered drugs that block muscular activity, conditioned responding may perfectly well occur to the conditioned stimulus when the paralytic drug has worn off (e.g. Solomon and Turner, 1962). Learning clearly occurs when the animals are paralysed, even though no behavioural changes take place at that time. The second problem is that in many cases it can be established that organisms do much more than simply acquire new types of behaviour. For instance, in a famous experiment, MacFarlane (1930) trained laboratory rats to run through a maze to obtain food, and found that when the maze was filled with water, the animals continued to obtain the food even though they now had to swim to reach it. Clearly, learning in this case does not merely involve the acquisition of a set of particular muscle activities conditioned to a set of stimuli, but instead involves acquiring knowledge of the spatial layout of the maze, with this knowledge capable of revealing itself in a variety of different ways.

For these reasons, a more cognitive view is that learning is an abstract term that describes a transition from one mental state to a second in which encoded information is in some way different. This transition may perfectly well take place without the development of any new behaviour, and furthermore may manifest itself in a variety of quite different behaviours. But although it avoids the problems associated with defining learning in terms of behaviour, this definition also has its shortcomings. For instance, how

are we to distinguish between learning and forgetting? Forgetting, like learning, can be viewed as a change in encoded information, except that in this case information is lost rather than gained. Our definition would plainly need to be supplemented by a proviso that learning involves the gain of information, but it is likely to be very difficult to specify what we mean by 'information gain' without relying simply on behaviour. We might find ourselves reverting to a behavioural definition of learning, which is precisely what we were trying to avoid.

Another problem with the cognitive definition is that it fails to deal satisfactorily with examples of what we might call 'non-cognitive' learning. The cognitive definition refers to a transition from one *mental* state to another, and the reason for incorporating the restriction to mental states is to exclude examples like the following. Roediger (1993) reports that the average duration of labour for first-born babies is about 9.5 hours, while that for later born babies is about 6.6 hours. Clearly, for second and third children the amount of time the mother spends in labour is much less than for first children. It seems strange to say that the female reproductive system is capable of 'learning' and 'remembering', so we would like to exclude this sort of case, despite the fact that information has obviously been acquired by the body. The restriction to mental states excludes the labour case because the relevant changes take place in the body without any mental component. But then we seem to be committed to saying that all habits and skills (which we do want to include as examples of learning) must be mental, and this seems unduly restrictive. Is it not likely that some aspects of learning a skill like playing tennis are really bodily rather than mental? Borderline cases like this probably illustrate the futility of trying to define learning.

Since the experimental study of learning began in the late nineteenth century, when Hermann Ebbinghaus (1885) commenced his pioneering laboratory investigations of human learning and evolutionists such as George Romanes (1882) began to use controlled experiments to investigate animal intelligence, the study of animal and human learning has continued in parallel, but regrettably not always with as much cross-reference as one might wish. Although in this book I focus solely on human learning, I hope that nothing concluded here will offend a student of animal learning. While humans may have learning capabilities that are available to few (or no) other organisms, such as the ability to abstract general rules (Mackintosh, 1988), I would argue that the correspondences between human and animal learning mechanisms far outweigh their differences. Some psychologists believe that since the motivation and prior experience of laboratory animals can be carefully controlled, it is research with animals that tends to produce the major discoveries about learning, with the study of human learning merely following along behind.

However, one of the principal aims of this book is to show that genuine

insights about learning have been made in the last decade or two of research on humans. Because amenable human subjects only require appropriate instructions in order to perform almost any task, no matter how bizarre, there are a number of things that can be investigated in humans that would be extremely difficult, if not impossible, to study in non-humans. Obvious examples include tasks requiring subjects to make similarity and probability judgements, from which a wealth of interesting findings have emerged. Further, data can be obtained from humans that are orders of magnitude more complex, and therefore theoretically challenging, than those obtainable from non-humans. The obvious example involves language acquisition, where even highly simplistic models need to be of great sophistication (e.g. Elman, 1990; Pinker, 1991). In addition to focusing solely on human learning, the discussion in this book is also restricted in that it will not extend to language learning. To cover language acquisition would of course require a book in itself, but I should mention that I believe that the ability to explain language learning is the touchstone of any theory of learning, and I would be surprised if language learning turns out to rely on mechanisms radically different from those discussed here.

Historical background

For much of the century following Ebbinghaus' (1885) pioneering studies of learning, research has been conducted either explicitly or implicitly within the associationist tradition. It is important to distinguish the term 'associative' as in 'associative learning' from the term 'associationism.' The former is a purely descriptive term referring to the type of learning that takes place – whatever its nature – when a relationship exists between certain events in the organism's environment. 'Associationism', in contrast, refers to a particular view of how that learning is effected: it is the view that in the final analysis, all knowledge is based on connections between ideas. Sensory systems provide an organism early in life with very simple perceptual experiences, which during development become associated as a result of their co-occurrence to yield more complex experiences. These associations are such that when one has been formed, it automatically carries the mind from one idea to another.

Associationism took a central place in the psychology of learning not only because of its simplicity but also, and more fundamentally, because it provided the bedrock for the empiricist analysis of the mind, and of science in general, which had become dominant by the nineteenth century. An obvious difficulty for the empiricist view that all knowledge is derived from experience is that many concepts or ideas, such as a biologist's concept of a gene, are infinitely more complex than the simple sensory experiences that, according to empiricism, provide the only foundation for knowledge. Associationism provides a potential solution in the hypothesis that primi-

tive experiences can become associated to yield more complex ideas, and those ideas can then in turn become associated to yield even more complex ideas, until all of the complexity of the biologist's concept is accounted for.

The associationist perspective provided both an overall conception of learning and knowledge, and also in the hands of Ebbinghaus and his followers the obvious means of investigating learning. If simple association of ideas is the only process involved, then all that is needed is to set up some simple associative learning task in which as many superfluous features as possible are removed, and use it to investigate the basic laws of learning. The learning of lists of nonsense syllables and of 'paired-associates' provided the ideal solution: learning that the nonsense syllable *wux* was on a list seems to require nothing more than the formation of associations between the phonetic elements of the syllable and associations between those elements and the list context, while learning the arbitrary response 'reason' to the stimulus word 'window' requires the formation of an associative bond between two pre-existing but previously unconnected ideas. Thus it was thought that laboratory studies of nonsense syllable and paired-associate learning would be sufficient to uncover *all* of the universal laws of learning.

By the 1950s the associationist analysis of learning had reached almost total dominance, to the point where many textbooks on learning and memory took it for granted that associationism provided the only explanatory framework worth considering. In the hands of researchers like Benton Underwood (1957) and Leo Postman (1962), paired-associate learning acquired the status in studies of human learning that the Pavlovian conditioning procedure has acquired in animal research and, as is clear from reviews such as that of Deese and Hulse (1967), sophisticated discussions of whether learning occurred gradually or was all-or-none, whether forgetting was due to trace decay or interference, seemed to imply that genuine progress was being made.

However, by the early 1970s cognitive psychologists had begun to tire not only of such artificial tasks as paired-associate learning, but also of learning in general. Partly, no doubt, this was due to the apparent intractability of some of the key issues: investigators had argued themselves to a standstill over the continuous versus all-or-none debate, for instance (Restle, 1965). But more important was a long-term shift of interest towards knowledge representation. Two particular strands to this shift are worth considering.

The associationist view that complex concepts can be reduced to the association of elementary ideas had been resisted throughout the first half of the century by a small minority of researchers, including Gestaltists such as Koffka (1935) and Kohler (1947). Rather than consisting of the association of ideas, the Gestaltists emphasised the importance of organisation, and viewed learning as the construction of an organised whole in which the associated items are subcomponents. On this view, learning does not pro-

ceed via the strengthening of simple connections between ideas: rather, it involves the construction of new entities, holistic memory traces representing the ideas, their conjunction, and the current context. Retrieval, similarly, does not involve the activation of one idea via the flow of energy along a connection, but rather the reactivation of an entire memory trace.

Moreover, it was not only the Gestaltists who took this view of learning. In a paper published in 1893, the philosopher James Ward asked the apparently simple question of why repetition improves memory, and challenged the typical associationist view that repetition leads to the gradual strengthening of a mental bond or connection. Instead, Ward proposed that each repetition lays down a quite separate memory trace, and that memory improves because more traces exist to be accessed. The largely-forgotten memory researcher Richard Semon also advocated such a multiple-trace view (see Schacter, Eich and Tulving, 1978), and combined it with sophisticated ideas about how retrieval occurs (he coined the term 'ecphory' that refers to the reactivation of a complete memory trace on presentation of a cue that matches part of it). The multiple-trace view has, of course, been continued in recent years both in Endel Tulving's (1983) work on memory retrieval and in Hintzman's (1976) use of frequency judgments to try to discriminate strength and multiple-trace views of repetition effects.

The holistic view of learning and representation gained support from a large number of animal discrimination–learning experiments conducted during the 1950s. Suppose an animal is shown two red stimuli on some trials and is rewarded for choosing the right-hand one, while on other trials, a pair of green stimuli is presented and reward is given for choosing the left-hand stimulus. Simple though this discrimination may be, it cannot be solved on the basis of associations between the simple co-occurring elements of the task (red, green, left, right, reward, non-reward). This is because each element should become equally associated with each other element: red and green, for instance, are equally associated with left and right and reward and non-reward. The fact that humans as well as laboratory rats can learn these discriminations (e.g. Bitterman, Tyler and Elam, 1955) argues that the simplest sort of associationist analysis is insufficient, although in Chapter 4 we will see that this sort of discrimination, which is called a *nonlinear* classification, can be dealt with by more modern associationist theories.

In contrast to the difficulty posed for associationism, Gestalt views of learning are sufficiently flexible to be untroubled by this sort of discrimination learning: the organism is assumed to memorise the entire set of elements occurring on a given trial, such as {red, right, reward}, with the elements merely being parts of a larger memory trace. The organism may now solve the discrimination when shown a red stimulus by recalling that *right* is the choice that has been rewarded {red, ?, reward}. This sort of analysis led Medin (1975) to formulate an explicit model of configurational learning in the Gestaltist tradition called the context model, and we shall see in Chapter

3 that this theory has had some striking successes in describing human learning data. The model adopts a radically different unit of analysis from that of traditional associationist accounts: instead of elements becoming associated with outcomes, it is memorised configurations or 'instances' that underlie learning. The overall degree of similarity between a test item and the ensemble of stored instances determines the response that the item evokes.

The second and perhaps more powerful reason for an increasing interest in knowledge representation arose from the advent of the computer as a new model of the mind which made associationism seem totally inadequate. With the development of knowledge-representation programming languages like Lisp, investigators such as Newell, Shaw and Simon (1958) quickly began to develop computer models of highly complex human abilities such as solving logic problems and playing chess. Not only must paired-associate learning have looked decidedly trivial in comparison, but also the explanatory apparatus of these new computational models was far richer than associationism allowed: languages like Lisp represent knowledge symbolically, which meant that inference was readily possible, and the success of these models clearly suggested that their symbolic data-structures corresponded to those that were actually used by the human mind. The idea that complex knowledge, as argued by Quillian (1968), consists of concepts connected in propositional networks by semantically-interpreted relations appears to be quite at variance with associationism.

The development of these richer views of knowledge representation had a concomitant influence on ideas about learning. If knowledge is represented propositionally, then learning must involve the construction of propositional structures via a set of pre-existing general symbol-manipulating procedures. Accordingly, as computer models evolved, it became increasingly popular to view learning as a form of hypothesis-testing or rule-induction, and detailed studies of rule-induction were carried out, most famously by Bruner, Goodnow and Austin (1956) and Hunt, Marin and Stone (1966). As we will see in Chapter 5, such an approach has continued to this day and offers some important insights into learning.

Not everyone was persuaded by the rule-induction view of learning, and dissatisfaction was greatest amongst researchers interested not so much in learning as in conceptual representation. In a seminal article, Eleanor Rosch (1973) pointed out that knowledge of everyday objects such as chairs and birds is unlikely to be based on inductively-learned rules, since such rules would have to specify certain necessary and sufficient features for an object to be a member of the category. Yet surely, she argued, no such features exist: what could the necessary and sufficient perceptual features possibly be that define the category *bird*? Rather than sharing a set of common defining features, each member of a category can be thought of as a set of features, with large degrees of overlap between the features of different members of the category but with none of the features being necessary or sufficient. On

this view, categories may show 'graded structure', with some members of the category possessing more of its characteristic features than others and hence being more typical. As Barsalou (1990) has shown, such graded structure is a property of almost all categories, a fact that encourages the view that categories are represented by mental prototypes which correspond to objects possessing all of the characteristic features of that category.

Prototype theories therefore suggest that the learning process involves abstracting the category prototype from the experienced exemplars. A novel item is classified according to its similarity to the prototype stimulus. We will consider the prototype approach in Chapter 3, but here it is worth briefly mentioning one historical development that played a major role in the construction and testing of the instance and prototype theories discussed in Chapter 3. Objects in the world vary on a large number of independent dimensions such as colour, size, height, and so on, and we can therefore represent each object as a point in a multi-dimensional physical space. Each object also corresponds to a point in a corresponding mental space, where the dimensions of the space are those that the perceptual system uses to represent stimuli. However, the physical and psychological spaces may not be at all similar. For example, colour is one of the most salient aspects of mental representation, but has no exact physical correlate (see Hardin, 1990) – it is a psychological property. In the late 1950s, researchers began to develop the tools needed to analyse psychological spaces. By obtaining *proximity* measures such as similarity ratings from all pairwise combinations of a set of stimuli, it is possible to recover the locations and organisation of the stimuli in psychological space using statistical procedures such as multi-dimensional scaling and cluster analysis (see Shepard, 1980). From the point of view of learning, these developments have had immense importance. As a consequence of a learning episode, some representation of the training items will be formed, perhaps a prototype. Responding to a test item will be a function of its similarity to this representation. But how do we know how similar it is? Techniques such as multi-dimensional scaling provide an answer and therefore allow us to predict with great accuracy how test items will be treated.

Thus research up to the mid-1970s had set the scene for a variety of alternatives to the traditional associationist theories. Healthy interest was being paid to prototype and instance theories and to rule induction processes. The subsequent development of each of these approaches will be the focus of Chapters 3 and 5, while Chapter 4 will discuss the various ways in which associationism has evolved over the last two decades.

Three questions about learning

In this brief historical review, I have discussed alternative theories of learning as if they are incompatible with one another and as if evidence in favour of one theory must necessarily weaken the support of the others. There is

no doubt that many researchers see the various theories in this way. Hintzman (1976), for example, was quite adamant that frequency judgment data supported the multiple-trace approach and disproved theories based on associative strength, while more recently Waldmann and Holyoak (1992) claimed that some data of theirs 'clearly refute connectionist learning theories that subscribe to an associationistic representation of events as cues and responses' (p. 233). But it is also possible to see these theories as differing in their level of analysis and thus as not necessarily incompatible. Perhaps processes which at one level of analysis are well-characterised as being associationistic can also be described, at another level, as involving prototype abstraction or instance memorisation?

Ever since Marr (1982) published his highly-influential analysis of levels of explanation, psychologists have had to consider quite carefully how their research questions should be framed. Marr distinguished very clearly between the questions of what the system is computing and how it is doing it. Such questions need to be answered in quite different ways; the 'what?' question cannot be answered, for instance, by citing some complex brain mechanism which is more appropriate for the 'how?' question. With respect to learning, the highest level requires us to consider what it is in the environment that the associative learning system is sensitive to, while lower levels concentrate on the internal characteristics of the learning mechanism itself.

Following Anderson's (1990) extensive discussion of the different sorts of questions that can be asked of the cognitive system, I shall adopt the view that theories of learning have to address the following three fundamental questions. The first asks whether associative learning is *normative* (and hence rational). An associative relationship consists of a temporal distribution of events, and normative theories tell us whether or not an objective relationship exists in a given situation. Such theories can therefore be regarded as providing independent measures of association. Although the framing of normative theories is perhaps more the business of philosophers and statisticians than of psychologists, we will see that consideration of such theories is highly relevant to an understanding of learning. If it turns out that people perform well in comparison with the norms provided by a rational analysis, then it is reasonable to conclude that some mental algorithm exists for computing the norms in question. Chapter 2 considers the view, recently developed in detail by Cheng and Holyoak (1994), that the appropriate normative theory of associative learning is *contingency* theory. This theory provides a means of determining for any given situation what the objective relationship is between a pair of events; on this theory, people behave normatively or rationally if they believe events to be related only when contingency theory specifies that they indeed are. By considering evidence that has accumulated during the last 20 years, we will ask whether the human associative learning system behaves in ways that would be judged 'rational' given the prescriptions of contingency theory.