Learning the general but not the specific

Amnesia patients have a normal ability to learn categories from examples, even though they fail to learn the examples themselves; computational models of brain function suggest how and why.

In a recent article in *Science* [1], Knowlton and Squire reported the striking finding that amnesia patients, despite their markedly deficient ability to remember the individual items they had recently seen, nevertheless have an unimpaired ability to learn the general characteristics of the categories represented by those items. After being presented with a set of items to learn, the amnesia patients were shown a set of test items and asked two questions about each: firstly, whether it was identical to any of the items in the original set, and secondly, whether it was in the same category as any of the items in the original set. Their performance on the first of these tasks was close to chance level, but they were as good as normal individuals at determining whether a test item belonged to the learned category or not.

At first glance this finding seems to be totally paradoxical. Most views that have been put forward about how we learn categories postulate that we store traces of individual items and then combine what we know about each into some more abstract or generalized trace. One view holds that abstraction occurs by comparing traces of individual items, storing common features in the generalized trace and discarding features that differ between traces. The new results of Knowlton and Squire strongly challenge this view, and raise the question of how we can generalize properly if we have not stored the individual items on which that generalization is based?

Knowlton and Squire's findings come within a scientific context that may shed some light upon the apparent paradox. First, there is an increasing understanding of the neural substrate of memory in the brain. It is now widely accepted, thanks to 20 years of work by Squire and others (reviewed in [2]), that the hippocampus and related structures in the medial temporal lobes of the brain play a crucial part in the formation of explicit 'declarative' memories for the contents of specific experiences. This work has further established that the hippocampus and related structures are totally unnecessary for another form of learning, usually called implicit or procedural learning. This is the form of learning that operates in tasks that do not require explicit recollection of a previous episode or event.

One task involving implicit learning is word-fragment completion [3], in which a subject is shown a list of words, such as 'window', and then is later shown a list of word fragments, such as 'win...', and is simply asked to complete the fragment with the first word that comes to mind that completes the fragment. Prior exposure to the word 'window' greatly increases the probability that the subject will complete the fragment 'win...' as 'window' rather than, say, 'winter' or some other word. Strikingly, this effect is independent of explicit recognition of the previous occurrence of the word 'window', and it is as strong in patients with severe hippocampal amnesia as in normal adults.

This implies that the processing of individual items leaves a detectable residue that is independent of the hippocampus. Other tasks indicate that these 'after-effects' can accumulate, resulting in the learning of a generalized skill. Mirror reading — reading text reflected in a mirror — is such a skill. Without practice, normal individuals and patients with hippocampal amnesia can read reflected text, but only very slowly; performance improves dramatically with practice, and the amnesia patients acquire the skill at a normal rate, even though in some cases they have no explicit recollection of the processing episodes during which the skill was acquired [4].

The second relevant aspect of the recent scientific context of Knowlton and Squire's work is the development of explicit computational models [5–8] that treat implicit learning as the result of small adjustments to the strengths of the synaptic connections between neurons. These models use simplified, neuron-like processing units together with associative learning rules — that is, rules that adjust the strengths of connections between neurons based on simultaneous presynaptic and postsynaptic neuronal activity. If, in such models, small adjustments to the synaptic connections are made in the course of each processing episode, then clear analogs of both fragment completion and skill learning are obtained. Processing the word 'window' increases the likelihood that an incomplete input will complete itself as 'window' because the changes to the connection strengths tend to cause neurons that represent some parts of the input to activate the neurons that represent the other parts [5]. When trained with examples of items embodying a consistent set of input–output relationships that define a skill, the network gradually learns the relationship [6].

Such model neural networks are also capable of learning to categorize from examples. During learning, a series of examples is presented. Each example excites activity in the network that can be considered as

consisting of two firing patterns, one representing the example itself and the other representing the example's category 'label'. For each example, the network makes connection adjustments that increase the likelihood of both firing patterns related to that example being activated. After exposure to a series of examples drawn from the same category, the connection changes accumulate and those due to idiosyncratic differences between the examples average out, so that accurate categorization of new examples is possible. Such examples have previously been used to simulate in some detail the results of category learning experiments with normal individuals [5,7].

Given this background knowledge about the neuropsychology of amnesia and the performance of simulated neural networks, we are now in a position to consider possible interpretations of Knowlton and Squire's results. One possibility is that removal of the hippocampus and related structures results in a reduction in the size of the changes in connection strengths that can occur elsewhere in the brain during learning. Simulation studies [8] show that model neural networks that make relatively small connection strength changes on each trial are clearly less sensitive to particular items, but after being presented with many items they actually respond slightly better to typical new examples. The reason for this is simply that, with smaller connection strength changes, the accumulated connection strengths are less affected by the idiosyncrasies of recent individual items and thus have, in some sense, a better representation of the general features that characterize the category. At intermediate levels of training comparable to those used by Knowlton and Squire, the difference between networks that make connection strength changes of different magnitudes is slight, so their finding of no difference in category learning ability between normal individuals and amnesia patients is consistent with the idea that the latter simply make smaller changes to their synaptic strengths in response to each item they are asked to learn.

One difficulty with this interpretation is that it fails to account for the fact that the rate of initial acquisition of a skill, such as mirror reading, is exactly the same in normal individuals and amnesia patients. However, the finding that smaller connection strength changes lead to better generalization in the long run provides a rationale for another possible interpretation of Knowlton and Squire's results [9]. Perhaps the human learning system is divided into two types of associative learning system, so that each can optimize its achievement of a different goal. One system, tied to the hippocampus and related structures in the medial temporal lobes, may be optimized for rapidly learning the contents of specific episodes and events, and the other, found widely throughout the neocortex outside of the hippocampal system, may be optimized for the discovery of the shared features needed for categorization and for the acquisition of cognitive skills. According to this view, the two systems are both necessary because the incompatibility of their goals: the extraction of common features requires slow learning through the gradual accumulation of small changes, whereas rapid learning of individual cases requires larger changes that tend to interfere with the discovery of common features [9–11].

In conclusion, Knowlton and Squire's results seem at first glance incompatible with common sense views about the basis of our ability to generalize from experience. But when interpreted in the context of recent advances, both in our understanding of the details of hippocampal amnesia and in the simulation of category learning in model neural networks, the findings make a great deal of sense. Damage to the hippocampus takes the system responsible for the acquisition of explicit memories out of commission, but leaves the system that discovers the shared features of events and experiences intact.

References

1. KNOWLTON BJ, SQUIRE LR: **The learning of categories: parallel brain systems for item memory and category knowledge.** *Science* 1993, 262:1747–1749.
2. SQUIRE LR: **Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans.** *Psychol Rev* 1992, 99:195–231.
3. GRAF P, SQUIRE LR, MANDLER G: **The information that amnesic patients do not forget.** *J Exp Psychol: Learn, Mem, Cogn* 1984, 10:164–178.
4. COHEN NJ, SQUIRE LR: **Retrograde amnesia and remote memory impairment.** *Neuropsychol* 1981, 19:337–356.
5. MCCLELLAND JL, RUMELHART DE: **Distributed memory and the representation of general and specific information.** *J Exp Psychol: Gen* 1985, 114:159–188.
6. RUMELHART DE, MCCLELLAND JL: **On learning the past tenses of English verbs.** In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume II.* Edited by McClelland JL, Rumelhart DE and the PDP research group. Cambridge, Massachusetts: MIT Press; 1986:216–271.
7. KNAPP A, ANDERSON JA: **A signal averaging model for concept formation.** *J Exp Psychol: Learn, Mem, Cogn* 1984, 10:616–637.
8. MCCLELLAND JL, RUMELHART DE: **Distributed memory and amnesia.** In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume II.* Edited by McClelland JL, Rumelhart DE and the PDP research group. Cambridge, Massachusetts: MIT Press; 1986:503–527.
9. MCCLELLAND JL, MCNAUGHTON BL, O'REILLY R, NADEL L: **Complementary roles of hippocampus and neocortex in learning and memory.** *Soc Neurosci Abstr 508.7* 1992 18:1216.
10. MCCLOSKEY M, COHEN NJ: **Catastrophic interference in connectionist networks: The sequential learning problem.** In *The psychology of Learning and Motivation: Advances in Research Theory.* Edited by Bower GH. New York: Academic Press; 1989, 24:109–165.
11. MCCLELLAND JL: **The organization of memory: a parallel distributed processing perspective.** *Revue Neurologique* (in press).

James L. McClelland, Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.