Learning optimal strategies in complex environments

Terrence J. Sejnowski

Salk Institute for Biological Studies, La Jolla, CA 92037; and Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92039

Behavior becomes difficult to analyze when there are many stimuli and many response options. As a consequence, in most laboratory experiments the numbers of stimuli and choices are limited, with the two-alternative forced-choice experiment the most widely adopted. This minimal approach has been successful in studying reinforcement learning, in which responses to rewarded stimuli lead to predictable changes in behavior (1). To what extent can the basic principles of reinforcement learning, coupled with a complex environment and a large memory, account for more complex behaviors? The leaders of the cognitive revolution in the 1950s assumed that reinforcement learning could not account for cognitive behaviors such as language and reasoning, but surprisingly, recent advances in computational theory and experimental studies have challenged this assumption. A tour de force study in PNAS (2) adds to this evidence by showing that reinforcement learning can explain not only behavioral choice in a complex environment, but also the evolution toward optimal behavior over a long time.

We make several eye movements every second when scanning a complex image, and the scan path is dramatically influenced by what we are thinking (3). In the study by Desrochers et al. (2), a monkey was free to scan an array of dots, one of which was randomly baited with a reward on each trial. After several sessions of learning, and without any instructions, the monkey quickly settled on a regular scan path that visited all of the dots once on each scan out of the infinite number of possible scan paths that the monkey could have adopted. Adopting a single scan path is a sensible solution to the problem of collecting the maximum reward over a fixed amount of time. However, not all regular paths were equally efficient in reaching the reward, and only one had a minimum cost in terms of distance traveled. Remarkably, over many trials and weeks of practice, the monkeys broke their initial habit and sequentially explored several other regular scan paths, gradually improving their efficiency, and one of the monkeys eventually found the unique optimum. This behavior is characteristic of systems that use stochastic gradient ascent to find better solutions, in contrast to the experimenters who found the optimal scan path by programming a computer to perform an exhaustive search through all possible regular scan paths.

It is not obvious that the behavior of these monkeys can be explained by reinforcement learning because the rewards are randomly placed, and there is a high degree of uncertainty in the sampling process. In such circumstances, whereby rewards are delayed and costs for each choice are learned by trial and error, reinforcement learning can in principle account for more complex behaviors. The key to getting reinforcement learning to solve a complex problem rapidly is to find a good representation of the state space that generalizes well and to have enough memory to represent the relative values of all possible actions. Brains have evolved all of the machinery needed to solve complex problems with reinforcement learning. Classical conditioning, the basic learning step in reinforcement learning, has been found in a wide range of species, including invertebrates, which suggests that it was an early innovation that evolved to cope with uncertain environments (5). Dopamine neurons in the brainstem predict future rewards consistent with temporal-difference reinforcement learning (6, 7). Neurons in the cortex reflect these reward predictions and are sensitive to trial-by-trial fluctuations (8), which could drive the exploration of different regular scan paths (2). Other domains where reinforcement learning has been found to be effective include birdsong learning (9) and finding the optimal Nash equilibrium in games played against an opponent (10). Finally, unsupervised learning, such as priming, continually improves the sensory and motor representations in the cortex, making them faster and more efficient (11). What is the range of strategies that this brain machinery can learn when confronted with a complex environment?

An impressive demonstration that reinforcement learning can solve difficult problems is TD-Gammon, a program that started as a beginner and improved by playing itself, eventually achieving world champion level of play in backgammon (12). Solely on the basis of the reward at the end of each game, TD-Gammon discovered new strategies that had eluded the best experts (Fig. 1). This illustrates the ability of reinforcement learning to solve the temporal credit assignment problem and learn complex strategies that lead to winning ways. Reinforcement learning has also had increasing success at playing another board game: Go, one of the most difficult games for humans to master. Last year, a computer program with a seven-stone handicap beat a 5 dan professional (13). Reinforcement learning has also been used to learn complex control laws. For example, flying a helicopter is much more difficult than flying an airplane, but a control system was trained with reinforcement learning to perform helicopter aerobatics (14).

Author contributions: T.J.S. wrote the paper. The author declares no conflict of interest. See companion article on page 20512.

E-mail: sejnowski@salk.edu.
Despite these successes the jury is still out on whether reinforcement learning can explain the highest levels of human achievement. Rather than add a radically new piece of machinery to the brain, such as a language module (15), nature may have tinkered with the existing brain machinery to make it more efficient.

Children have a remarkable ability to learn through imitation and shared attention (16), which might greatly speed up reinforcement learning by focusing learning on important stimuli. We are also exceptional at waiting for rewards farther into the future than other species, in some cases delaying gratification to an imagined afterlife made concrete by words. Supercharged with a larger cerebral cortex, faster learning, and a longer time horizon, is it possible that we solve complex problems in mathematics the same way that monkeys find optimal scan paths?

ACKNOWLEDGMENTS. My laboratory is supported by the Howard Hughes Medical Institute.