Learning optimal strategies in complex environments

Terrence J. Sejnowski[1]
Salk Institute for Biological Studies, La Jolla, CA 92037; and Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093

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

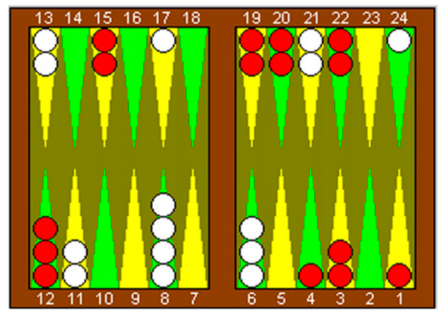

Fig. 1. TD-Gammon makes a brilliant move. This backgammon board confronted Joe Sylvester, then the highest-rated player in the world, in the championship match of the 1988 World Cup of Backgammon tournament. Sylvester, playing White, had rolled 4-4 and played 8-4*, 8-4, 11-7, 11-7. TD-Gammon's surprising recommendation is 8-4*, 8-4, 21-17, 21-17. Traditional human thinking would reject this play, because the 21 point was thought to be a better defensive anchor than the 17 point, and the 7 point a better blocking point than the 11 point. Analysis confirmed that TD-gammon's choice was far superior, and as a consequence many experts have now revised their evaluation of complex positional battles (12).

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 find the optimal scan path (4). However, the number of trials required to find the optimal path grows rapidly with the number of possible choices, so it is not clear how long it would take. A simple statistical reinforcement algorithm, however, replicated the monkey's behavior in remarkable detail, including the sequential search through regular scan patterns, in the same number of trials that the monkey took (2). Thus, even relatively unconstrained behavioral tasks can now be studied with the same rigor as forced-choice tasks.

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.
[1]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.

1. Glimcher PW, Camerer C, Poldrack RA, Fehr E, eds (2009) *Neuroeconomics: Decision Making and the Brain* (Academic Press, New York).
2. Desrochers TM, Jin DZ, Goodman ND, Graybiel AM (2010) Optimal habits can develop spontaneously through sensitivity to local cost. *Proc Natl Acad Sci USA* 107:20512–20517.
3. Yarbus AL (1967) *Eye Movements and Vision* (Plenum, New York).
4. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
5. Hammer M, Menzel R (1995) Learning and memory in the honeybee. *J Neurosci* 15:1617–1630.
6. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
7. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936–1947.
8. Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304:1782–1787.
9. Doya K, Sejnowski TJ (1995) A novel reinforcement model of birdsong vocalization learning. *Advances in Neural Information Processing Systems 7*, eds Tesauro G, Touretzky DS, Leen T (MIT Press, Cambridge, MA), pp 101–108.
10. Dorris MC, Glimcher PW (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* 44:365–378.
11. Moldakarimov SB, Bazhenov M, Sejnowski TJ (2010) Representation sharpening can explain perceptual priming. *Neural Comput* 22:1312–1332.
12. Tesauro G (1995) Temporal difference learning and TD-Gammon. *Commun ACM* 38:58–68.
13. Blincoe R (April 30, 2009) Go, going, gone? *The Guardian*, http://www.guardian.co.uk/technology/2009/apr/30/games-software-mogo.
14. Abbeel P, Coates A, Ng AY (2010) Autonomous helicopter aerobatics through apprenticeship learning. *Int J Robot Res*, 10.1177/0278364910371999.
15. Fodor J (1983) *The Modularity of Mind* (MIT Press, Cambridge, MA).
16. Meltzoff AN, Kuhl PK, Movellan J, Sejnowski TJ (2009) Foundations for a new science of learning. *Science* 325:284–288.