Dopamine Made You Do It

Terrence Sejnowski

DOPAMINE-RELEASING NEURONS are a core brain system that controls motivation.[1] When enough dopamine-releasing neurons die, the symptoms of Parkinson's disease appear; these include motor tremor, difficulty initiating actions, and, eventually, anhedonia, the complete loss of pleasure in any activity. The end stage includes catatonia, a complete lack of movement and responsiveness. But when the dopamine neurons are behaving normally, they provide brief bursts of dopamine to the neocortex and other brain areas when an unexpected pleasure (reward) occurs (be it food, money, social approval, or a number of other things) and a diminution of activity when less than expected reward is experienced (this can be a smaller reward or no reward at all).

Your dopamine neurons can be polled when you need to make a decision. What should I order from the menu? You imagine each item, and your dopamine cells provide an estimate of the expected reward. Should I marry this person? Your dopamine cells will give you a gut opinion that is more trustworthy than reasoning. Problems with many different dimensions are the most difficult to decide. How do you trade off a sense of humor in a mate, a good dimension, against being messy, a bad dimension, or hundreds of other comparisons? Your brain's reward systems reduce all these dimensions down to a common currency, the transient dopamine signal. Dopamine neurons receive inputs from a part of

the brain called the basal ganglia, which in turn receive input from the entire cerebral cortex. The basal ganglia evaluate cortical states and are involved with learning sequences of motor actions to achieve a goal.

The dark side of reward is that all addictive drugs act by increasing the level of dopamine activity. In essence, drugs like cocaine and heroin (as well as nicotine and alcohol) hijack the dopamine reward system, making your brain believe that taking the drug is your most important immediate goal. Withdrawal symptoms dominate when drugs are not immediately available. This motivates desperate actions to obtain more drugs, and such actions can jeopardize life and livelihood. Even after an arduous rehabilitation process, which can take years, the brain's reward circuit is still altered by the experience of addiction, leaving the recovering addict vulnerable to a relapse. This can be triggered by people and places, sounds and smells previously associated with the drugs, or even paraphernalia used to take the drugs. For an addict, dopamine is deeply compelling.

The basal ganglia are part of all vertebrate brains. Within the basal ganglia the dopamine neurons mediate a form of learning called associative learning, made famous by Pavlov's dog. In Pavlov's experiment, a sensory stimulus such as a bell (a conditioned stimulus) was followed by the presentation of food (an unconditioned stimulus), which elicited salivation even without the bell (an unconditioned response). After several pairings, the bell by itself would lead to salivation (a conditioned response). Different species have different preferred stimuli to associate. Bees are very good at associating the smell, color, and shape of a flower with the rewarding nectar, and they use this learned association to find similar flowers that are in season. Something about this universal form of learning must be important, and there was a period in the 1960s when psychologists intensively studied the conditions that gave rise to associative learning and developed models to explain it.

Only the stimulus that occurs just before the reward becomes associated with the reward.[2] This makes sense since the stimulus is more likely to have caused the reward if it comes before the reward than a stimulus just after the reward. Causality is an important principle in nature.

Suppose you have to make a series of decisions to reach a goal. If you don't have all the information about the outcomes of the choices ahead of time, you have to learn as you make the choices in real time. When

you get a reward after a sequence of decisions, how do you know which of the several choices you made were responsible? A learning algorithm that can resolve this issue, called the temporal credit assignment problem, was discovered by Richard Sutton at the University of Massachusetts at Amherst in 1988.[3] He had been working closely with Andrew Barto, his thesis adviser, on difficult problems in reinforcement learning, a branch of machine learning inspired by associative learning in animals. In temporal difference learning, you compare your expected reward for making a particular choice with the actual reward you get and change your expected reward so that next time you will be able make a better decision. Then an update is made to the value network that computes the future expected reward for each decision at each choice point. The temporal difference algorithm converges to the optimal series of decisions after you have had enough time to explore the possibilities. This is followed by a period of exploiting the best strategy found during the exploration.

Bees are champion learners in the insect world. It takes only a few visits to a rewarding flower for a bee to remember the flower. This fast learning was being studied in the laboratory of Randolph Menzel in Berlin when I visited him in 1992. The bee brain has around a million tiny neurons, and it is very difficult to record their electrical signals because they are so tiny. Martin Hammer in Menzel's group had discovered a unique neuron, called VUMmx1, that responded to sucrose (a type of sugar) with electrical activity but not to an odor; however, after the odor was delivered, followed shortly by the sucrose reward, VUMmx1 would now respond to the odor.

When I returned to La Jolla, Peter Dayan, a postdoctoral fellow in my lab who was an expert on reinforcement learning, immediately realized that this neuron could be used to implement temporal difference learning. Our model of bee learning could explain some subtle aspects of bee psychology, such as risk aversion. For example, when a bee is given a choice between a constant reward and twice the amount but at 50 percent probability (on average the same amount), bees prefer the constant reward. Read Montague, another postdoctoral fellow in my lab, took the next leap and realized that dopamine neurons in the vertebrate reward system may have a similar role in our brains.[4] In one of the most exciting scientific periods of my life, these models and their predictions were published and subsequently confirmed in monkeys with single neuron

recordings by Wolfram Schultz and in humans with brain imaging.[5] Transient changes in the activity of dopamine neurons signal reward prediction error.

Temporal difference learning might seem weakly effective since the only feedback present is whether or not you are rewarded at the end of a sequence of actions. However, several applications of temporal difference learning have shown that it can be powerful when coupled with other learning algorithms. Gerry Tesauro worked with me on the problem of teaching a neural network to play backgammon. Backgammon is a highly popular game in the Middle East, and some make a living playing high-stakes games. It is a race to the finish between two players, with pieces that move forward based on each roll of the dice, passing through each other on the way. Unlike chess, which is deterministic, the uncertainty with every roll of the dice makes it more difficult to predict the outcome of a particular move. The knowledge of backgammon in Gerry's program was captured by a value function that provided an estimate of winning the match from all possible board positions as ranked by a panel of backgammon experts. A good move can be found simply by evaluating all possible moves from the current position and choosing the one with the highest value.

Our approach used expert supervision to train neural networks to evaluate game positions and possible moves. The flaw in this approach is that many expert evaluations of board positions were needed and the program could never get better than our experts. When Gerry moved to the IBM Thomas J. Watson Research Center, he switched from supervised learning to temporal difference learning and had his backgammon program play itself. The problem with self-play is that the only learning signal is a win or a loss at the end of the game with no information about the contribution of the many individual intermediate moves during the game to that win or loss.

At the beginning of the backgammon learning, the machine's moves were random, but eventually one side won. The reward first taught the program how to "bear off" and exit all of the pieces from the board at the end of the game. Once the endgame was learned, the value function for bearing off in turn trained the value function for the crucial middle game, where subtle decisions need to be made about engagements with the other

player's pieces. Finally, after playing a hundred thousand games, the value function was honed to play the opening, in which pieces take defensive positions to prevent the other player from moving forward. Learning proceeds from the end of the game, where there is an explicit reward, back toward the beginning of the game, using the implicit reward learned by the value function. What this shows is that by back-chaining with a value function, it is possible for a weak learning signal like the dopamine reward system to learn a sequence of decisions to achieve a long-term goal.

Tesauro's program, called TD-Gammon, surprised me and many others when he revealed it to the world in 1992.[6] The value function had a few hundred model neurons in it, a relatively small neural network by today's standards. After a hundred thousand games, the program was beating Gerry, so he alerted Bill Robertie, an expert on positional play in backgammon from New York City, who visited IBM to play TD-Gammon. Robertie won the majority of games but was surprised to lose several well-played games and declared it the best backgammon program he had ever played. Several of the moves were unusual ones that he had never seen before; on closer examination these proved to be improvements on typical human play. Robertie returned when the program had reached a million self-played games and was astonished when TD-Gammon played him to a draw. A million may seem like a lot, but keep in mind that after a million games, the program saw only an infinitesimal fraction of all possible board positions. Thus TD-Gammon was required to generalize to new board positions on almost every move.

In March 2016, Lee Sedol, the Korean Go World Champion, played a match with AlphaGo, a program that learned how to play Go using temporal difference learning.[7] AlphaGo used neural networks with a much larger value network, having millions of units to evaluate board positions and possible moves. Go is to chess in difficulty as chess is to checkers. Even Deep Mind, the company that had developed AlphaGo, did not know its strength. AlphaGo had played hundreds of millions of games with itself, and there was no way to benchmark how good it was. It came as a shock to many when AlphaGo won the first three games of the match, exhibiting an unexpectedly high level of play. Some of the moves made by AlphaGo were revolutionary. AlphaGo far exceeded what I and many others thought was possible. The convergence between biological intelli-

gence and artificial intelligence is accelerating, and we can expect even more surprises ahead. The lesson we have learned is that nature is more clever than we are.

We are just beginning to appreciate the powerful impact of dopamine on making decisions and guiding our lives. Since the influence of dopamine is subconscious, the story we tell ourselves to explain a decision is probably based on experiences no longer remembered. We make up stories because we need to have conscious explanations. Every once in a while we have a "gut feeling" about a choice that does not have an easy explanation—it was the dopamine that made us do it.

NOTES

1. E. Bromberg-Martin, M. Matsumoto, O. Hikosaka, "Dopamine in Motivational Control: Rewarding, Aversive, and Alerting," *Neuron* 68 (2010): 815–834.
2. There are some notable exceptions to the notion that the conditioned stimulus must immediately precede the unconditioned stimulus in associative learning. One is food-aversion learning. If you eat something and then become ill hours later, you will still strongly associate that food with the illness and tend to avoid that food in the future, even though the conditioned stimuli (the sight, smell, and taste of the food) can precede the unconditioned stimulus (feeling ill) by several hours.
3. R. S. Sutton, "Learning to Predict by the Method of Temporal Differences," *Machine Learning* 3 (1988): 9–44.
4. P. R. Montague, P. Dayan, and T. J. Sejnowski, "A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning," *Journal of Neuroscience* 16 (1996): 1936–1947.
5. W. Schultz, P. Dayan, and P. R. Montague, "A Neural Substrate of Prediction and Reward," *Science* 275 (1997): 1593–1599.
6. G. Tesauro, "Temporal Difference Learning and TD-Gammon," *Communications of the ACM* 38 (1995): 58–68.
7. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (2016): 484–489.