

Lecture 14: Dopaminergic value learning

1. A reward signal is regarded as ‘pure’ if it doesn’t reflect specific facts about the sensory properties of the reward cue or expectation, and if it doesn’t vary with the type of action appropriate to harvesting the reward. The majority of midbrain dopamine neurons are ‘almost pure’ reward signalers in this sense. They fall short of ‘true’ purity because they send signals that prepare the motor system for one or another member of a set of broad action types.
2. Why should we expect to find purity or near-purity? A mobile animal must make many quick decisions that involve choices between, and therefore comparisons of values amongst, different kinds of contingencies (food, mating opportunities, care for young, exploration of territory). This requires polysensory processing and a common internal currency. *Here is the real source of the economics in neuroeconomics.*
3. Latencies of dopamine neurons are < 100 ms, and phasic firing durations are < 200 ms.
4. The error response of dopamine neurons varies with the reward value, and isn’t sensitive to different combinations of magnitude and probability that combine to yield identical values. Value is operationalized by about 65% of the neurons as mean divided by the standard deviation; thus the same level of activation will occur to the highest mean-valued reward within any given distribution. The remaining 35% of neurons are increasingly active as variance increases; thus these neurons can code for risk levels. The relevant sampling window is < 2 seconds.

5. A stimulus that is paired with a fully predicted reward does not become a reward predictor. Such a stimulus is said to be *blocked*. No dopamine depression will follow its presentation without the reward, and dopamine activation will follow reward delivery after it is presented alone.
6. Dopamine neurons aren't found only in the midbrain, though that is the only part of the brain they (overwhelmingly) dominate. There are also dopamine neurons in OFC. These are less pure reward signalers, because they distinguish between more specific kinds of rewards, and prepare more specific kinds of motor responses.
7. Along with signaling information about reward values and risk, dopamine neurons also set thresholds that must be met for various motor, cognitive, motivational and learning processes to function properly. This makes it difficult to interpret behavioral consequences of changes in midbrain dopamine levels: some will be products of learning and some will be consequences of changes in global 'readiness' for attention and action.
8. A first experiment: humans under fMRI learned that light flashes predicted squirts of liquid. No interesting differences were observed upon receipt of juice compared with water. But much of interest was seen in responses to the predictor cues. The following conditions were varied: (i) predictability of the reward type (juice or water); (ii) predictability of the reward timing; (iii) whether subjects were active or passive. The authors don't report results based on variation (i). However, variation (iii) made no difference, indicating that the dopamine response isn't only coding for decisions to act. Subjects exhibited strong hemodynamic variations in striatum in accordance with the hypothesis that their brains were predicting the expected times of squirts. This was

important in light of a long-established result from conditioning literature that animals learn more efficiently (i.e., remember for longer) when stimulus-response intervals vary.

9. Timing prediction seems pointless unless it is entirely for the sake of priming the motor system – *or* unless the brain can assess the value of maintaining its attention and processing by comparing the promise of its predictor system with the value of some alternative (perhaps a default). Berns and Montague develop a model that represents the valuation of reliance on the output of a predictor.
10. The predictor valuation (PV) model is characterized as follows. Suppose $R(x,n)$ estimates the value of a reward distributed at various possible times x, y, z, \dots, n in the future, scaled according to the uncertainty attending to the intervals between the estimation point and each time, as in:

$$R(x,n; D) = \int_{-\infty}^{+\infty} dy G(x - y, (x - n)D)r(y)$$

where $G(z,b) = (2\pi b)^{-1/2} \exp\{-z^2/2b\}$ and D is a constant. Then the value $F(n)$ the brain attaches to getting a particular predictor signal at perceptual time n is given by:

$$F(n) = \int_n^{+\infty} dx e^{-q(x-n)} \int_{-\infty}^{+\infty} dy G(x - y, (x - n)D)\rho(y) = \int_n^{+\infty} dx \{e^{-q(x-n)}\} \times \{R(x,n; D) = \int_n^{+\infty} dx \{discount\ future\ time\ x\ relative\ to\ perceptual\ time\ n\} \times \{diffused\ version\ of\ reward\ estimate\ \rho(x)\ for\ some\ x\ and\ n\}$$

11. The diffusion term reflects the idea that as time to reward increases, the probability of error increases with it.
12. Berns and Montague speculate, based on fMRI data from one (!) monkey, that PV might be

implemented in OFC and striatum. (This speculation isn't daring. Given what you've learned about striatum, if PV really is computed in the brain, where else would you go looking for it first?)

13. The model predicts foraging patterns of honeybees who must allocate their relative investments between high-yield, high-variance flowers and lower-variance flowers with the same EV. They allocate about 20% of their time to the risky flowers and 80% to the safe ones.
14. The model does not predict optimization when learning spaces include local maxima. (See the right-hand graph in Figure 11 of Berns and Montague's article.) In an experiment with a group of human subjects given a task with such a pattern, subjects split into two roughly equal-sized groups. One group became stuck at local maxima as predicted. However, the other group played through reward troughs and maximized utility. Similar behavior has subsequently been observed, in two separate experiments, in populations of monkeys. If this sort of heterogeneity governs investment choices in natural populations, we would expect the risk-tolerant maximizers to get steadily richer relative to the conservative matchers.
15. The authors found that strength of correlation of activity in left NAcc with changes in predictability of rewards predicts risk-tolerant maximization in people. This result was later found in monkeys under single cell recording.
16. Though derived entirely independently from it, the PV model is structurally isomorphic to the Black-

Scholes model of pricing and hedging risky asset portfolios in efficient markets. Berns & Montague suggest that this isn't a coincidence. The Black-Scholes model, though intended as normative, is supposed to advise investors on what to do in markets that behave as real historical markets (of the relevant kind) have done. B&M speculate that perhaps these markets behaved as they did because the participants in them were using human brains, running the PV algorithm, to make their decisions.

17. There is a specific basis for skepticism about this hypothesis. As B&M say, the PV model predicts hyperbolic discounting. But it isn't true, as is often claimed, that the best interpretation of the empirical data is that most people hyperbolically discount. This is the result one seems to get if one assumes that the discounting of everyone in a (suitably large) sample must be modeled using the same function. Then the hyperbolic function will beat its rivals because some people are strongly attracted to immediate rewards. But what predicts best of all are mixture models that allow for heterogeneous discount functions, with some subjects – typically the greatest number – modeled as discounting exponentially, at least following a short front-end delay before the earliest reward options. (See the *Econometrica* paper by Andersen *et al* cited in a previous lecture.) We would especially expect to observe exponential discounting in financial asset markets with sophisticated traders. Note that mixture models control for risk attitudes, on which data should be obtained independently, so the heterogeneity is not explained away in the manner B&M attempt, by hypothesizing a subset of risk-tolerant or risk-loving

agents. (It's true that one typically finds such subsets, however.)