Does phasic dopamine signalling play a causal role in reinforcement learning?

Peter Shizgal*
Centre for Studies in Behavioural Neurobiology
Department of Psychology
Concordia University
Montréal, QC H4B 1R6
Canada
peter.shizgal@concordia.ca

Ivan Trujillo-Pisanty
Center for the Neurobiology of Addiction, Pain, and Emotion
Department of Anesthesiology and Pain Medicine
Department of Pharmacology
University of Washington
Seattle, WA 98195
ivantp@uw.edu

Marie-Pierre Cossette
Centre for Studies in Behavioural Neurobiology
Department of Psychology
Concordia University
Montréal, QC H4B 1R6
Canada
mpy_cossette@hotmail.com

Kent Conover
Centre for Studies in Behavioural Neurobiology
Department of Psychology
Concordia University
Montréal, QC H4B 1R6
Canada
kentlconover@gmail.com

Francis Carter
Centre for Studies in Behavioural Neurobiology
Department of Psychology
Concordia University
Montréal, QC H4B 1R6
Canada
francis44carter@gmail.com

Vasilis Pallikaras
Centre for Studies in Behavioural Neurobiology
Department of Psychology
Concordia University
Montréal, QC H4B 1R6
Canada
vpallikaras@gmail.com

Yannick-André Breton
Caprion Biosciences
Montréal, QC H2X 3Y7
yannick.breton@gmail.com

Rebecca Brana Solomon
Centre for Studies in Behavioural Neurobiology
Department of Psychology
Concordia University
Montréal, QC H4B 1R6
Canada
rb_solomon@hotmail.com

Abstract

The reward-prediction-error hypothesis holds that payoff from future actions can be maximized and reward predictions optimized by incremental adjustment of connection weights in neural networks underlying expectation and choice. These adjustments are driven by reward prediction errors, discrepancies between the experienced and expected reward. Phasic firing in midbrain dopamine neurons is posited to both represent reward-prediction errors and to cause the weight changes these errors induce. There is abundant correlational evidence from rodents, monkeys, and humans that midbrain dopamine neurons encode reward-prediction errors. The work discussed here tests and challenges the causal component of the reward-prediction-error hypothesis of dopamine activity. Rats were trained to self-administer rewarding electrical stimulation of the medial forebrain bundle or optical stimulation of midbrain dopamine neurons. Stimulation-induced release of dopamine was monitored by means of fast-scan cyclic voltammetry. Both forms of stimulation triggered reliable, recurrent release of dopamine in the nucleus accumbens. According to the RPE-DA hypothesis, such repeated,

*David Munro built and maintained the computer-controlled equipment for experimental control and data acquisition. Software for experimental control and data acquisition was written and maintained by Steve Cabilio. PS web site: http://www.concordia.ca/research/neuroscience/faculty.html?fpid=peter-shizgal. Video re ICSS and the measurement of reward intensity: https://spectrum.library.concordia.ca/978205/

response-contingent release should eventually drive action weights into saturation. If unopposed by a countervailing influence, the repeated release of dopamine should render stable reward-seeking performance at non-maximal levels impossible. Instead, the rats performed at stable non-maximal levels in response to intermediate stimulation strengths.

Keywords: brain stimulation reward; intracranial self-stimulation; medial forebrain bundle; ventral tegmental area; nucleus accumbens; electrical brain stimulation; optogenetics; fast-scan cyclic voltammetry

Acknowledgements

The authors are grateful to Peter Dayan, Ritwik Niyogi, and Sanjeevan Ahilan for many fruitful discussions of the issues addressed here, for their original modeling work on other aspects of the phenomena, and for sharing their insights and interpretations. This work was supported by a Natural Sciences and Engineering Research Council of Canada grant (RGPIN-2016-06703) to PS. The TH-Cre rats used in the optogenetic experiments were sourced from a colony estabilshed with founders kindly donated by Ilana Witten and Karl Deisseroth.

Reward-prediction errors, dopamine, and intracranial self-stimulation

An elegant, enormously influential hypothesis about the nature and neural mechanisms of learning holds that reward-prediction errors (RPEs), encoded in the firing of dopamine (DA) neurons, optimize expectations about future rewards and the values assigned to reward-seeking actions [1]. The seminal paper introducing this reward-prediction-error hypothesis of dopamine neuron activity (RPE-DA hypothesis) [1] applies temporal-difference reinforcement-learning (TDRL) methods [2] within an actor-critic framework [3].

Paramount in the paper by Montague et al. [1] is their incisive account of DA activity in monkeys performing tasks that combine Pavlovian and operant conditioning. Also included is a brief discussion of how the TDRL model applies to electrical intracranial self-stimulation (eICSS), the performance of instrumental tasks to trigger activation of brain circuitry. The authors point out that DA cell bodies in the ventral tegmental area give rise to axons that course through eICSS sites along the medial forebrain bundle (MFB). Fig. 1 summarizes their portrayal of eICSS, which holds that stimulation of eICSS sites produces a fictive RPE by activating DA neurons.

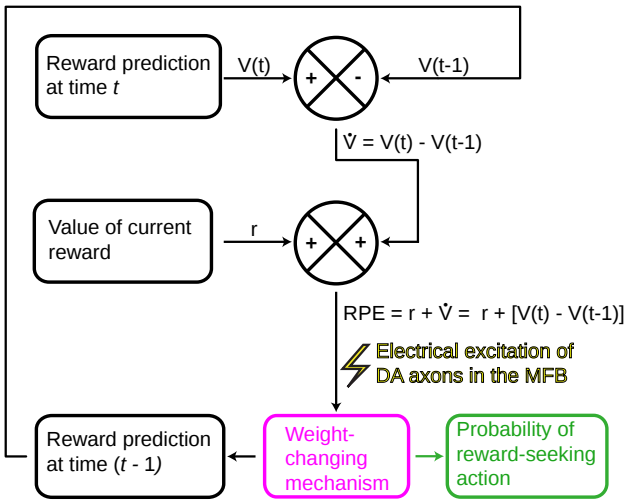

Figure 1: Portrayal of eICSS in [1]

Optogenetic methods make it possible to activate midbrain DA neurons exclusively, unlike electrical stimulation, which is less selective. Rodents will work for such optical stimulation (oICSS) [4, 5, 6]. Specific optical activation of midbrain DA neurons has also been shown to augment responding to a redundant reward-predicting cue that would otherwise have been behaviorally ineffective and to delay extinction of responding to a cue no longer paired with the delivery of a sucrose reward [7]. These results were interpreted as evidence for a causal role of DA-mediated RPEs in learning.

Here, we summarize eICSS and oICSS experiments that reassess the causal role of DA-mediated RPEs in learning. The behavioral and electrochemical findings are not easily explained by the RPE-DA hypothesis.

Two problems with the original TDRL portrayal and a potential remedy

We note two problems with the portrayal in Fig. 1 as it applies to eICSS of the most extensively studied stimulation site: the lateral hypothalamic (LH) level of the MFB. First, the positioning of the electrode isolates the RPE from its corrective consequences. In the case of natural rewards earned under stable conditions, the RPE "predicts itself away." The RPE renders the prediction progressively more accurate, and hence the RPE shrinks progressively to zero. In contrast, an RPE elicited by means of direct axonal stimulation of DA axons in the MFB would be of constant magnitude because it arises beyond the regions where the signals encoding the predicted ($V_t - V_{t-1}$) and experienced (r) rewards must be combined: the somatodendritic region of the DA neurons and/or their afferent network. (Recall that under appropriate conditions, DA firing is perturbed in opposite directions by these two signals.) If the DA neurons were activated downstream from the point(s) at which these two input signals converge, both the reward prediction and the weight assigned to the reward-seeking action (e.g., lever pressing) would be driven over repeated iterations to their maximal (saturated) values. Stable performance for electrically induced rewards of intermediate magnitude would thus be impossible. This prediction is contradicted by abundant evidence from operant matching experiments in which the value of intermediate-strength rewards is stable over many repeated reward encounters, e.g. [8].

The consequences of weight saturation due to stimulation of DA axons are illustrated in panel A of Fig. 2. The agent earns rewards by performing a sustained action ("work"), such as depressing a lever for a required duration. Multiple rewards can be earned during a trial. When the reward is weak ($r = 1$), the latency to begin working is long. However, due to the unconditional RPE, the action weight is boosted by delivery of each reward. Thus, the cumulative work-time accelerates until it attains its maximal velocity. The stronger the reward ($r_5 > r_4 > r_3 > r_2 > r_1$), the shorter the latency to reinitiate responding after reward delivery and the faster the growth of the action weight. Panel B shows the contrasting case of a natural reward.The RPE shrinks as the prediction improves, and the action weights stabilize at values proportional to the reward strength. Thus, the slope of the cumulative work-time trajectory is scaled by the reward strength.

A second problem with the schema in Fig. 1 is the implicit attribution of the behavior to direct activation of DA axons. These small-diameter axons are unmyelinated and have very high thresholds to activation by extracellular currents [9]. Moreover, psychophysical estimates of conduction velocity, recovery from refractoriness, and frequency following in the directly activated MFB fibers subserving eICSS of the MFB implicate neurons with myelinated axons much more readily excited than those of the DA neurons [10].

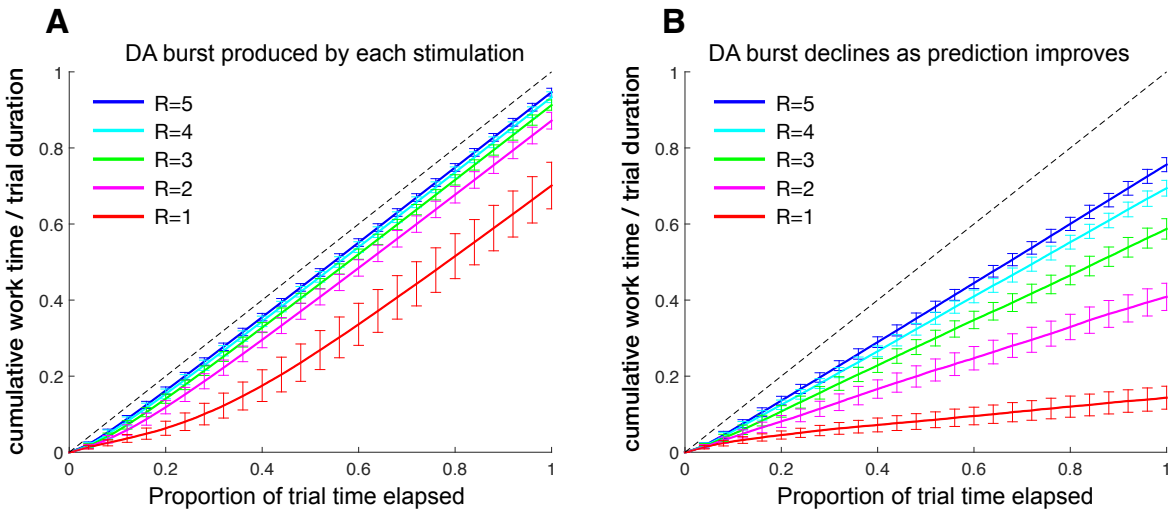

Figure 2: Simulations by Peter Dayan

Both problems could be solved by changing the assumption about where the stimulation intervenes in the TDRL schema (Fig. 3). The rewarding effects of MFB stimulation and intraoral sucrose compete and summate, suggesting that electrical stimulation of the MFB mimics the value of a natural reward [11]. If the electrode indeed excites neurons subserving the primary reward signal (r) instead of those subserving the RPE, then the temporal-difference signal ($V_t - V_{t-1}$) could come to null this electrically induced input, and trajectories such as those shown in panel B of Fig. 2 could be achieved. However, as we show below, this remedy is unavailable in the case of oICSS, which entails direct, unconditional activation of DA neurons. Thus, Fig. 3 predicts: a) oICSS trajectories like those in panel A of Fig. 2 coupled to continued simulation-induced DA release, but b) eICSS trajectories like those in panel B of Fig. 2 coupled to decline and cessation of simulation-induced DA release. We tested these predictions.

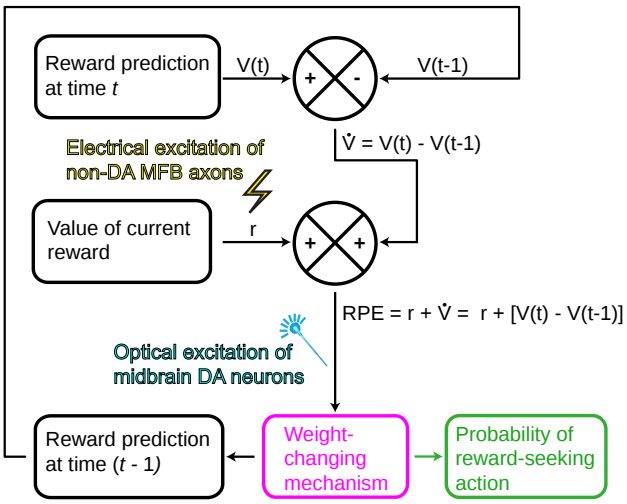

Figure 3: Revised TDRL account of eICSS and oICSS

Methods

ICSS. Electrodes for eICSS were aimed at the LH level of the MFB. Stimulation consisted of 0.5 s trains of constant-current pulses, 0.1 ms in duration. To prepare TH-Cre(+/-) rats for oICSS, channelrhodopsin-2 (ChR2) was expressed in midbrain DA neurons via Cre-Lox recombination and viral transfection, and 300 μm-core optical fibers were aimed at the ventral tegmental area (VTA). Optical stimulation consisted of 1 s trains of 462 or 473 nm pulses, 5 ms in duration.

The triadic-trial paradigm. Experimental sessions were comprised of trials arranged in cycling triads. During the leading trial of each triad, the strength (pulse frequency) of the stimulation was set to the maximum the rat could tolerate, whereas during the trailing trial, it was set to a negligibly rewarding value. The stimulation strength on offer during each central ("test") trial of the triads was also constant within a trial, but it varied across triads, and was selected at random from a vector 3-14 elements in length.The maximum and minimum values of the vector were the strengths used in the leading and trailing trials, respectively.

Electrochemistry. The extracellular DA concentration was measured by means of fast-scan cyclic voltammetry (FSCV). Carbon-fiber microsensors were aimed at the nucleus accumbens (NAc), and an Ag/AgCl reference electrode was positioned 10.7 mm caudal to the NAc. Cyclic voltammograms were generated at 10 Hz by applying an 8.5 ms triangular waveform that ramped from -0.4 V to $+1.3$ V and back to -0.4 V at a scan rate of 400 V/s. A modification of the method of Kishida et al. [12] was used to extract DA concentrations: principal-component regression was substituted for elastic-net regression.

Results and discussion

Fig. 4 shows empirical data from one rat, averaged over 19 sessions, from test trials during which the stimulation strength was sampled randomly from a 14-element vector. Cumulative work time rises at a roughly constant rate, which depends systematically on the strength of the rewarding stimulation. In 22 rats performing eICSS, we obtained 29 datasets

using test-trial stimulation-strength vectors 9 or 14 elements in length. Like those in Fig. 4, the slopes of the cumulative work-time trajectories vary systematically as a function of stimulation strength and are linear or mildly concave downwards. In no case do the data resemble the simulated results in panel A of Fig. 2, which show initial acceleration towards a constant terminal slope due to the unconditional RPE. Instead, the results are consistent with panel B of Fig. 2, in which terminal slopes are related systematically to reward strength, and with the revised TDRL schema in Fig. 3. In the revised schema, stable performance at intermediate levels is achieved because the TD signal can null the input from the stimulation electrode, thus eliminating the RPE and the firing of DA neurons that encodes it. To find out whether DA release indeed ceases during stable eICSS performance at intermediate levels of performance, we measured DA release in the NAc during eICSS by means of FSCV.

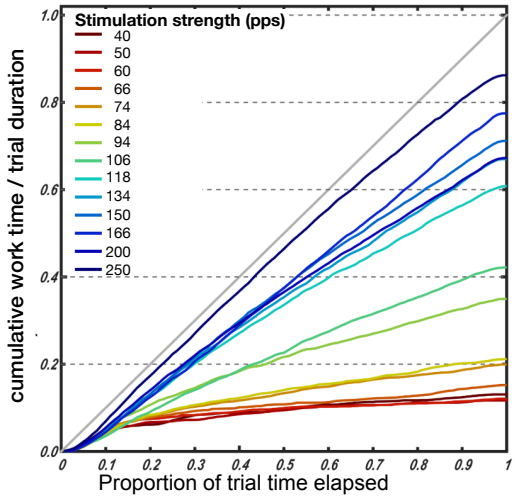

Figure 4: eICSS: stable performance for intermediate-strength electrical rewards

The FSCV recordings were obtained while the rat performed eICSS in a simplified version of the triadic-trial paradigm. Only three stimulation strengths were sampled on test trials: the High and Low values were the same as on leading and trailing trials, respectively, whereas the Med value was intermediate. Behavioral data from one rat, averaged over two test sessions, are shown in panel B of Fig. 5. Again, performance for the medium-strength (Med) reward is roughly stable over the course of the trial. Panel A shows the corresponding measurements of DA concentration, which are cumulations of the peak post-stimulation DA concentration measured following delivery of each stimulation train (panel A of Fig. 6). According to the RPE-DA hypothesis, roughly stable performance at intermediate work levels can be achieved only in the absence of persistent, recurring DA-mediated RPEs. However, panel A of Figs. 5 and 6 show that stimulation-induced DA release continued throughout the eICSS trial, thus calling the hypothesis into question.

Panel B of Fig. 6 shows that optical stimulation of midbrain DA neurons, like electrical stimulation of the MFB (panel A), persistently and reliably elicits transient increases in DA concentration. According to the RPE-DA hypothesis, such transients should alter action weights in the manner depicted in panel A of Fig. 2: when an intermediate-strength reward is on offer during the test trial, the cumulative work trajectory should accelerate until it achieves the maximum slope that the rat's physical capacity allows. This is not what we found.

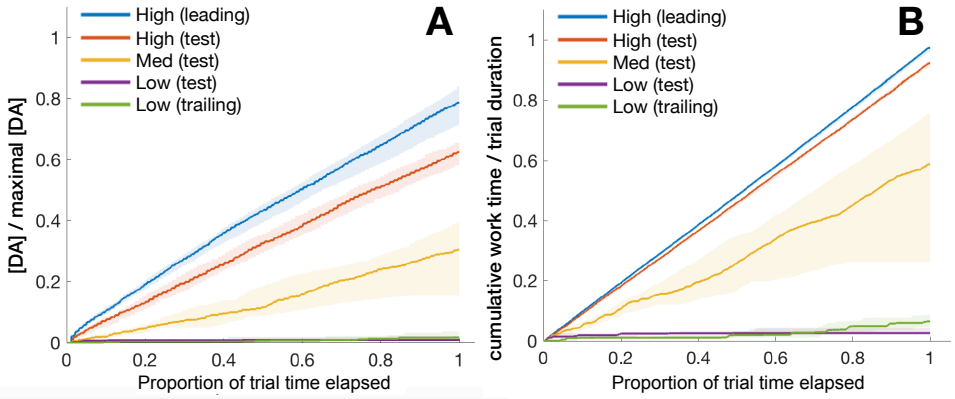

Figure 5: Concurrently acquired FSCV and behavioral data

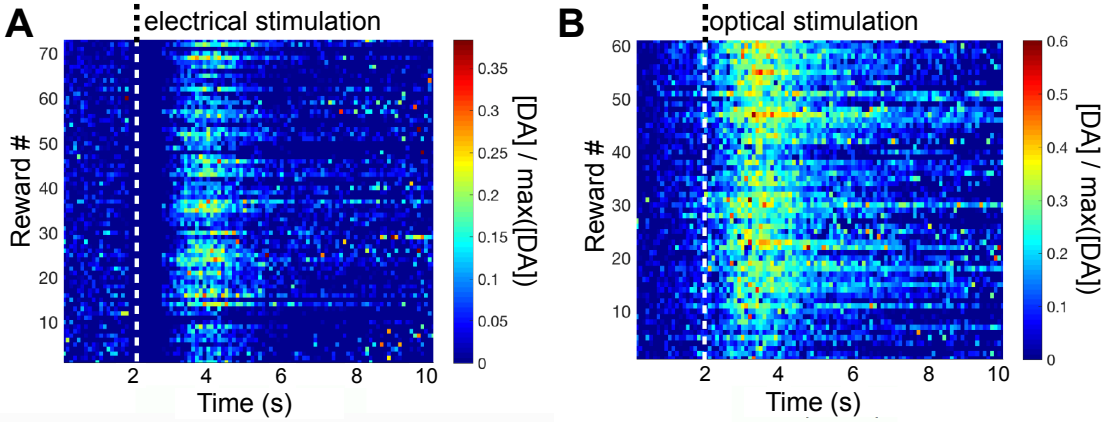

Figure 6: DA transients driven by electrical or optical stimulation

Fig. 7 shows data from a rat working for optical stimulation of midbrain DA neurons in the simplified triadic-trial paradigm. When the reward strength was intermediate, a stable, linear work trajectory is observed. Linear or slightly concave-downward trajectories were also shown by five additional rats performing oICSS in the simplified triadic-trial paradigm. Like the eICSS data, these results call into question a key aspect of the RPE-DA hypothesis, the notion that DA-mediated RPEs cause changes in action weights.

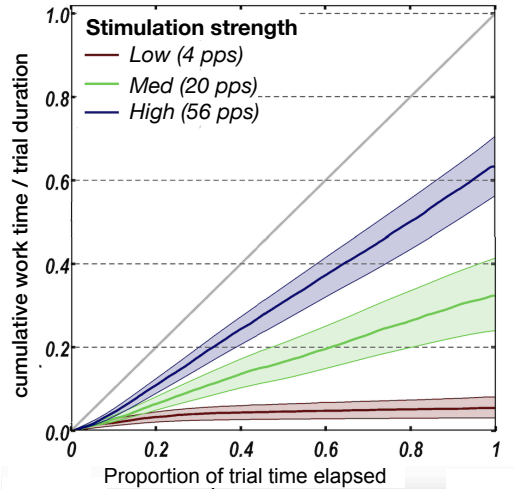

Figure 7: Stable performance for medium-strength optical activation of DA neurons

Reconciling the results with the RPE-DA hypothesis. Peter Dayan has proposed a way to reconcile the present findings with the RPE-DA hypothesis. Could reward predictions come to decrease DA firing in unstimulated neurons, thus compensating for the excitation of the subpopulation of DA neurons recruited by the stimulation? The low baseline firing rate (3-5 spikes s^{-1}) of DA neurons poses a problem for this proposal: the baseline is much closer to zero than to the maximum firing rate. Thus, inhibition of multiple unstimulated DA neurons could be required to compensate for the excitation of each stimulated neuron. This would be difficult to achieve given the massive, bilateral recruitment of midbrain DA neurons by electrical MFB stimulation. That said, this proposal merits rigorous experimental test.

Limitations. In the different versions of the triadic-trial paradigm, stable behavioral data has been obtained from 26 rats performing working for electrical stimulation and 9 rats working for optical stimulation. However, concurrent measurements of behavior and DA concentration have been carried out successively in only two rats to date. Additional subjects must be tested, and the FSCV recording sites must be adjusted in the light of recent findings showing functional specialization of NAc subregions [13].

Conclusion. The findings reported here raise serious questions about the causal component of the RPE-DA hypothesis and provide several proofs of principle for novel ways to test this foundational idea.

References

[1] P. R. Montague, P. Dayan, & T. J. Sejnowski, "A framework for mesencephalic dopamine systems based on predictive Hebbian learning", *The Journal of neuroscience*, vol. 16, no. 5, pp. 1936–1947, 1996.

[2] R. S. Sutton, "Learning to predict by the methods of temporal differences", *Machine Learning*, vol. 3, no. 1, pp. 9–44, 1988.

[3] J. C. Houk, J. L. Adams, & A. G. Barto, "A model of how the basal ganglia generate and use neural signals that predict reinforcement", in J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the Basal Ganglia*, pp. 249–270, The MIT Press, 1995.

[4] A. R. Adamantidis, H.-C. Tsai, B. Boutrel, *et al.*, "Optogenetic Interrogation of Dopaminergic Modulation of the Multiple Phases of Reward-Seeking Behavior.", *The Journal of neuroscience*, vol. 31, no. 30, pp. 10829–10835, 2011.

[5] C. D. Fiorillo, "Transient activation of midbrain dopamine neurons by reward risk", *Neuroscience*, vol. 197, no. C, pp. 162–171, 2011.

[6] I. B. Witten, E. E. Steinberg, T. J. Davidson, *et al.*, "Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement.", *Neuron*, vol. 72, no. 5, pp. 721–733, 2011.

[7] E. E. Steinberg, R. Keiflin, J. R. Boivin, *et al.*, "A causal link between prediction errors, dopamine neurons and learning", *Nature Neuroscience*, vol. 16, pp. 966–973, 2013.

[8] M. I. Leon, V. Rodriguez-Barrera, & A. Amaya, "The effect of scopolamine on matching behavior and the estimation of relative reward magnitude.", *Behavioral Neuroscience*, vol. 131, no. 5, pp. 406–420, 2017.

[9] J. S. Yeomans, N. T. Maidment, & B. S. Bunney, "Excitability properties of medial forebrain bundle axons of A9 and A10 dopamine cells.", *Brain research*, vol. 450, no. 1-2, pp. 86–93, 1988.

[10] P. Shizgal, "Neural basis of utility estimation", *Current Opinion in Neurobiology*, vol. 7, no. 2, pp. 198–208, 1997.

[11] K. L. Conover & P. Shizgal, "Competition and summation between rewarding effects of sucrose and lateral hypothalamic stimulation in the rat", *Behavioral Neuroscience*, vol. 108, no. 3, pp. 537–548, 1994.

[12] K. T. Kishida, I. Saez, T. Lohrenz, *et al.*, "Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward", *Proceedings of the National Academy of Sciences*, vol. 113, no. 1, pp. 200–205, 2016.

[13] J. W. de Jong, S. A. Afjei, I. Pollak Dorocic, *et al.*, "A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System", *Neuron*, vol. 101, no. 1, pp. 133–151.e7, 2019.