Value representations do not explain movement selectivity in DMS-projecting dopamine neurons

Rachel S. Lee, Marcelo G. Mattar, Nathan F. Parker, Ilana B. Witten, Nathaniel D. Daw

Princeton Neuroscience Institute
Department of Psychology
Princeton University, Princeton, NJ, USA 08550

Abstract

Dopaminergic (DA) neurons in the midbrain provide topographic innervation of the striatum that is essential for learning and generating movements. Although traditionally DA neurons have been thought to primarily encode reward prediction error (RPE), recent studies have found movement-related signals specifically in DA neurons that project to the dorsal striatum. However, whether these movement signals should still be best understood in the traditional framework as a specialized RPE with respect to the movement in question remains a major open question. For example, we recently reported that DA neurons that project to the dorsomedial striatum (DMS) are primarily modulated by choices contralateral to the recording site. Is this modulation by contralateral choice in fact a contralateral movement signal, or a contralaterized RPE signal? Here, we resolve this question by examining DA responses while carefully considering both choice and RPE on a trial-by-trial basis. We show that DA responses are modulated by contralateral choice with a pattern that is qualitatively distinct from RPE. This implies that choice encoding cannot be explained by RPE and is better explained by the direction of movement. These results demonstrate a fundamental separation in RPE and movement encoding, which may help shed light on the diversity of functions and dysfunctions of the DA system.

Introduction

A central feature of dopamine (DA) is its association with two apparently distinct functions: reward and movement (Niv et al. 2007; Berke 2018). Although manipulation of DA produces gross effects on movement initiation and invigoration, physiological recordings of DA neurons have historically shown few neural correlates of motor events (Wise 2004; Schultz, Dayan, and Montague 1997). Instead, classic studies reported responses to rewards and reward-predicting cues, with a pattern suggesting that DA neurons carry a "reward prediction error" (RPE) – the difference between expected reward and observed reward – for learning to anticipate rewards (Schultz, Dayan, and Montague 1997; Andrew G. Barto 1995; Cohen et al. 2012; Coddington and Dudman 2018; Soares, Atallah, and Paton 2016). In this classic framework, rather than explicitly encoding movement, DA neurons influence movements indirectly, by determining which movements are learned, and/or the general motivation to engage in a movement (Niv et al. 2007; Collins and Frank 2014; Berke 2018).

However, complicating this classic view, several recent studies have suggested that subpopulations of DA neurons may instead have a more direct role in encoding movements. For example, we recently reported that whereas dopamine neurons projecting to ventral striatum showed classic RPE responses, a subset of midbrain DA neurons that project to the dorsomedial striatum (DMS) were selective for a mouse's choice of action. In particular, they responded more strongly during contralateral (versus ipsilateral) choices as mice perform a probabilistic learning task (Parker et al. 2016). In addition, there have been several other recent studies that reported phasic changes in DA activity at the onset of spontaneous movements (Dodson et al. 2016; M. W. Howe and Dombeck 2016; da Silva et al. 2018; Barter et al. 2015).

These recent observations leave open an important question: can the putatively movement-related responses be reconciled with Reinforcement Learning (RL) models describing the classic RPE response? One possibility is that movement-related responses also reflect RPEs, but for reward predictions tied to particular movements. Specifically, computational models like the actor-critic (A. G. Barto, Sutton, and Anderson 1983) and advantage learning (Baird 1994) learn separate predictions about the overall value of situations or stimuli and about the value of specific actions. It has long been suggested these two calculations might be localized to ventral vs dorsal striatum, respectively (Montague, Dayan, and Sejnowski 1996; O'Doherty et al. 2004; Takahashi, Schoenbaum, and Niv 2008). Furthermore, a human neuroimaging experiment reported evidence of distinct prediction errors for right and left movements in the corresponding contralateral striatum (Gershman, Pesaran, and Daw 2009).

This leads to the specific hypothesis that putative movement-related signals in dorsal-projecting DA neurons might actually reflect an RPE related to the predicted value of contralateral choices. A choice-specific RPE could explain choice correlates, because temporal difference RPEs do

not do not just signal error when a reward is received, they also have a phasic anticipatory component triggered by predictive cues indicating the availability and timing of future reward, such as (in choice tasks) the presentation of levers or choice targets (Montague, Dayan, and Sejnowski 1996; Morris et al. 2006; Roesch, Calu, and Schoenbaum 2007). This anticipatory prediction error is proportional to the value of that future expected reward– indeed, we henceforth refer to this component of the RPE as a "value" signal, which tracks the reward expected for a choice. Crucially, a choice-specific value signal can masquerade as a choice signal because, by definition, action and value are closely related to each other: animals are more likely to choose actions they predict have high value–so, for instance, a value signal (RPE) for the contralateral choice will tend to be larger when that action is chosen than when it is not (Samuelson 1938). Altogether, given the fundamental correlation between actions and predicted value, a careful examination of the neural representation of both quantities is required to determine whether or not movement signals can be better explained as value-related.

Thus, to address this possibility, we sought to closely examine neural correlates of value and movement in our DA recordings in mice performing a probabilistic learning task. Since value predictions are subjective, we estimated value in two ways: 1) by using reward on the previous trial as a simple, theory-neutral proxy, and 2) by fitting the behavioral data with a more elaborate trial-by-trial *Q*-learning model. We compared the observed DA modulations to predictions based on modulation either by movement direction, and/or the expected value (anticipatory RPE) of contralateral or chosen actions.

Ultimately, although we detected previously unappreciated value-related modulation at the time of lever presentation in both the terminals and cell bodies of DMS-projecting DA neurons, the modulation reflected the value of the chosen action rather than the contralateral one. Thus, value could not explain contralateral choice selectivity in these neurons, implying that this choice-dependent modulation in fact reflects modulation by contralateral movements and not value.

Results

Task, behavior and DA recordings

As described previously, mice were trained on a probabilistic reversal learning task (Parker et al. 2016). Briefly, each trial began with an illumination in the nose port, which cued the mice to initiate a nose poke (**Figure 1a**). After a 0-1s delay, two levers appeared on both sides of the nose port. Each lever led to reward either with high probability (70%) or low probability (10%), with the identity of the high probability lever swapping in a pseudorandom schedule after a block of at least 10 rewarded trials (**Figure 1b**). After another 0-1s delay, the mice either received a sucrose reward and an accompanying auditory stimulus (positive conditioned stimulus, or CS+), or no reward and a different auditory stimulus (negative conditioned stimulus, or CS-).

Given that block transitions were not signaled to the mouse, after each transition mice gradually learned to prefer the lever with the higher chance of reward. To capture this learning, we fit their choices using a standard trial-by-trial *Q*-learning model that predicted the probability of the animal's choice at each trial of the task (**Figure 1c**). In the model, these choices are driven by a pair of decision variables (known as *Q*-values) putatively reflecting the animal's valuation of each option.

As mice performed this task, we recorded activity from either the terminals or cell bodies of DA neurons that project to DMS (VTA/SN::DMS) using fiber photometry to measure the fluorescence of the calcium indicator GCaMP6f (**Figure 1d,e; Supplemental Figure 1a,b**). As previously reported, this revealed elevated activity during contralateral choice trials relative to ipsilateral choice trials, particularly in relation to the nose poke and lever presentation events (**Figure 1f,g; Supplemental Figure 1c**) (Parker et al. 2016).

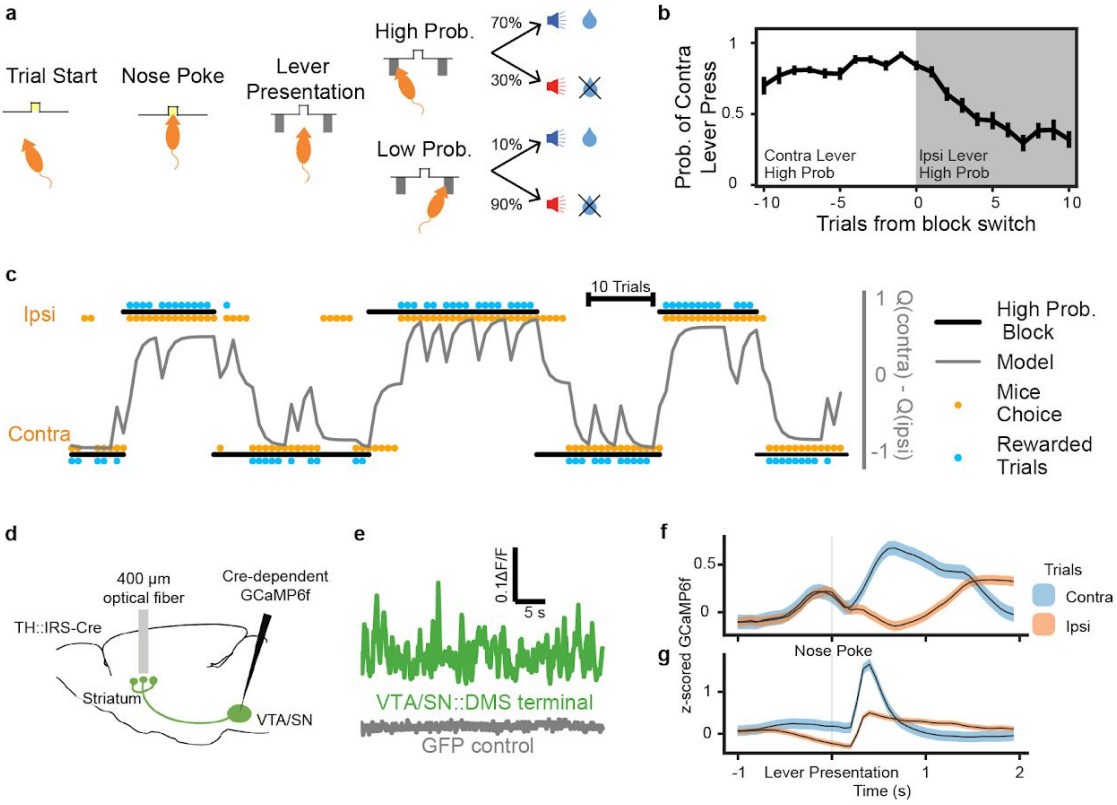


Figure 1: Mice performed a probabilistic reversal learning task during GCaMP6f recordings from VTA-SN::DMS terminals or cell bodies. (**a**) The illumination of the central nosepoke signaled the start of the trial, allowing the mice to enter the nose port. After a 0-1 second jitter delay, two levers are presented to the mice, one of which results in a reward with high probability (70%) and the other with a low probability (10%). The levers swapped probabilities on a pseudorandom schedule, unsignaled to the mouse. (**b**) The averaged probability of how likely the mice are to choose each lever before and after the identity of the high probability switched. Error bars indicate 1

standard error. “Contra” and “Ipsi” refer to the location of the lever relative to the side of the recording. (**c**) We fit behavior with a trial-by-trial Q learning mixed effect model. Example trace of 200 trials of a mouse's behavior compared to the model results. Black bars above and below the plot indicate which lever had the high probability for reward; Orange dots indicate mice’s actual choice; Blue dots indicate whether or not mice was rewarded; Grey line indicate the difference of the model’s Q values for contralateral and ipsilateral choices. **d**) Surgical schematic for recording with optical fibers from the GCaMP6f terminals originating from VTA-SN. Projections were determined using viral traces. (**e**) Sample GCaMP6f traces from VTA/SN::DMS terminals and aGFP control animal. **(f, g)** Previous work has reported contralateral choice selectivity in DMS DA terminals (Parker et al. 2016) when the signals are time-locked to nose poke (**f**) and lever presentation (**g**).

Predictions of Contralateral and Chosen Value Models

We introduce two hypothetical frames of reference by which the DMS DA neurons activity may be modulated by predicted value during trial events prior to the outcome: the DA responses could be modulated by the value of the contralateral option (relative to ipsilateral; **Figure 2a**), or by the value of the chosen option (relative to unchosen; **Figure 2b**). Note that both of these can be understood as the anticipatory component (occasioned at lever presentation) of a temporal difference RPE, with respect to the respective action’s value.

The first possibility is modulation by the value of the contralateral (relative to ipsilateral) action (**Figure 2a**; such responses have been reported in human neuroimaging, Gershman et al., 2009, Palmenteri et al. 2009; but not previously to our knowledge examined in dopamine unit recordings in animals). Assuming mice tend to choose the option they expect to deliver more reward, such responses would be larger, on average, during contralateral choices than ipsilateral ones (**Figure 2a**). Thus, when the DA responses are broken down with respect to both the action chosen, and its value, the direction of value modulation would depend on the choice: responses are highest for contralateral choices when these are relatively most valuable, but the lowest for ipsilateral choices when *they* are most valuable.

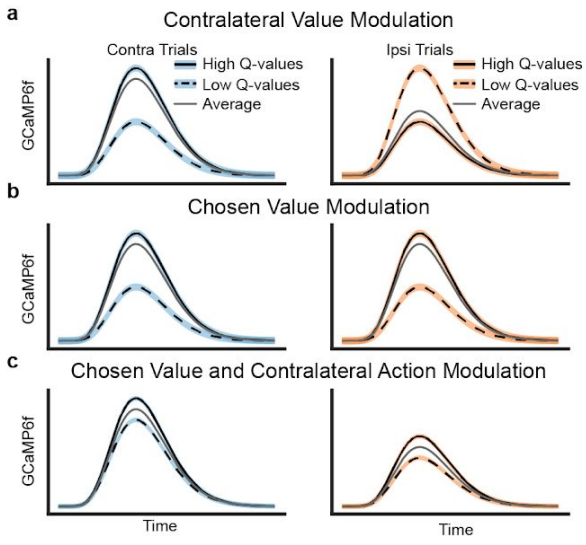


Figure 2: Schematics of possible types of value modulation at lever presentation. Trials here are divided based on Q values of chosen minus unchosen action. **(a)** Contralateral value modulation theory postulates that the responses are selective for the *value* of the contralateral action (relative to ipsilateral value) instead of the action itself. This means that the direction of value modulation should be flipped for contralateral versus ipsilateral choices. Since mice would more often choose an option when its value is higher, the average GCaMP6f responses would be higher for contralateral than ipsilateral choices. **(b)** Alternatively, the responses may be modulated by the value of the chosen action, resulting in similar value modulation for contralateral and ipsilateral choice. This type of value modulation will not in itself produce contralateral selectivity seen in previous results. (**c**) However, if the responses

were modulated by the chosen value and the contralateral choice, the averaged GCaMP6f would exhibit the previously seen contralateral selectivity.

The second possibility is that value modulation is relative to the chosen (versus unchosen) option (**Figure 2b**). Indeed, human neuroimaging studies have primarily reported correlates of the value of the chosen option in dopamine target areas (Daw et al., 2006; Boorman et al., 2009; Li & Daw, 2011) and also has been observed in primate dopamine neurons (Morris et al., 2006).

If DMS-projecting DA neurons indeed display chosen value modulation (**Figure 2b**), rather than contralateral value modulation, the value modulation for both contralateral and ipsilateral choices would be similar. Therefore, value modulation could not in itself account for preferential responding to contralateral actions that we observed in these neurons. In this case, to account for contralateral choice preference, one would have to assume DA neurons are also selective for the contralateral action itself (unrelated to their value modulation; **Figure 2c**).

DA in dorsomedial striatum is modulated by chosen value, not contralateral value

In order to determine which type of value modulation better captured the signal in DA neurons that project to DMS, we compared the GCaMP6f signal in these neurons for high and low value trials. We focused on the lever presentation since this event displayed a clear contralateral preference (**Figure 1g**). As a simple and objective proxy for the value of each action (i.e., the component of the RPE at lever presentation for each action), we compared responses when the animal was rewarded (high value), or not (low value), on the previous trial. (To simplify interpretation of this comparison, we included only trials in which the mice made the same choice as the preceding trial, which accounted for 76.6% of the trials.) The traces (**Figure 3a**) indicated that the VTA/SN::DMS terminals were modulated by the previous trial's reward. The value-related responses reflected chosen value – higher responding when the previous choice was rewarded, whether contralateral or ipsilateral – and therefore did not explain the movement-related effect. This indicates that the DMS-projecting DA neurons represent both chosen value and movement direction during the trial (similar to **Figure 2c**).

We repeated this analysis using trial-by-trial Q values extracted from the model, which we reasoned should reflect a finer grained (though more assumption-laden) estimate of the action's value. (For this analysis, we were able to include both stay and switch trials.) Binning trials by chosen (minus unchosen) value, a similar movement effect and value gradient emerged as we had seen with the previous trial outcome analysis (**Figure 3b**). Trials with higher Q values had larger GCaMP6f responses, regardless which side was chosen, again suggesting that VTA/SN::DMS terminals were modulated by the expected value of the chosen (not contralateral) action, in addition to being modulated by contralateral movement.

To quantify these effects statistically, we used a linear mixed effects regression at each of time point of the time-locked GCaMP6f. The explanatory variables included the action chosen (contra or ipsi), the differential Q values (oriented in the reference frame suggested by the data, chosen minus unchosen), the value by action interaction, and an intercept (**Figure 3c**). The results verify significant effects of both movement direction and action value; that is, although a significant value effect is seen, it does not explain away the movement effect. Furthermore, the appearance of a consistent chosen value effect across both ipsilateral and contralateral choices is reflected in a significant value effect and no significant interaction during the period when action and value coding are most prominent (0.25 - 1 seconds after lever presentation), as would have been predicted by the contralateral value model. (There is a small interaction between the variables earlier in the trial, before 0.25 seconds, reflecting small differences in the magnitude of value modulation on contralateral versus ipsilateral trials.) Conversely, when the regression is re-estimated in terms of contralateral value rather than chosen value, a sustained, significant interaction does emerge, providing formal statistical support for the chosen value model; see **Supplemental Figure 2.**

We performed the same value modulation analyses on the cell bodies, rather than terminals, of VTA/SN::DMS neurons (**Figure 3d-f**). This was motivated by the possibility that there may be changes in neural coding between DA cell bodies and terminals due to synapses forming directly on DA terminals. In this case, we found very similar modulation by both chosen value and contralateral movement in both recording locations.

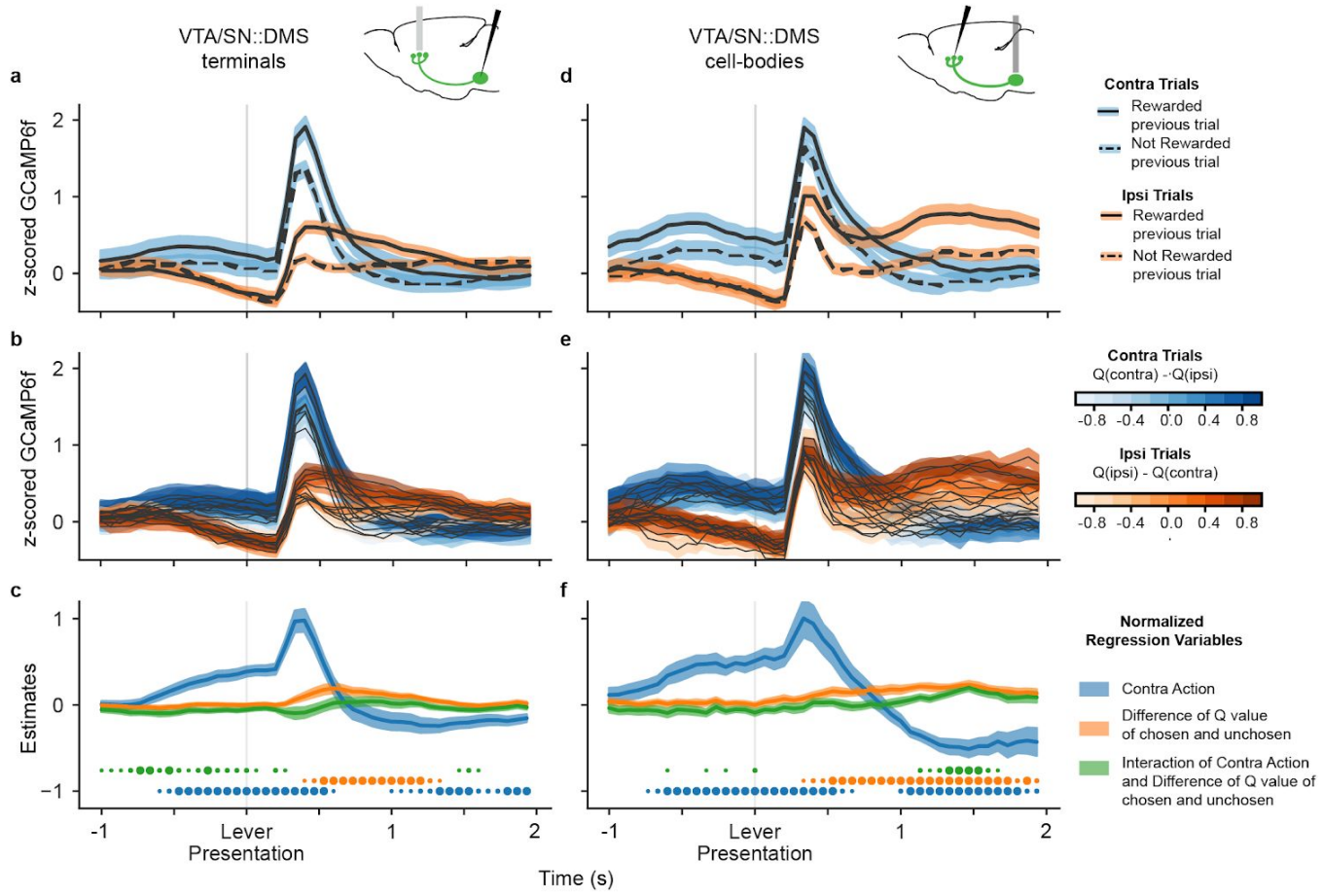


Figure 3: DA neurons that project to DMS are modulated by both chosen value and movement direction. (a) GCaMP6f signal relative to the lever presentation time for contralateral trials (blue) and ipsilateral trials (orange), as well as rewarded (solid) and non-rewarded previous trial (dotted) from VTA/SN::DMS terminals. Colored fringes represent 1 standard error from activity averaged across recording sites (n = 12). **(b)** GCaMP6f signal for contralateral trials (blue) and ipsilateral trials (orange), and further binned by the difference of Q values of chosen and unchosen action. Colored fringes represent 1 standard error from activity averaged across recording sites (n = 12). **(c)** Mixed effect model regression on each datapoint from 3 seconds of GCaMP6f traces. Explanatory variables include the action of the mice (blue), the difference in Q values for chosen vs unchosen actions (orange), their interaction (green), and an intercept. Colored fringes represent 1 standard error from estimates. Dots at bottom mark timepoints when the corresponding effect is significantly different from zero at p<.05 (small dot), p<.01 (medium dot), p<.001 (large dot). **(d-f)** Same as **(a-e),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

Direction of movement predicts DMS DA responses

An additional observation supports the interpretation that the contralateral choice selectivity in DMS-projecting DA neurons is related to the direction of movement, and not the value of the choice. When the responses are time-locked to the lever press itself, there is a reversal of the response selectivity between contralateral and ipsilateral trials, shortly after the lever press

(**Figure 4a, b**). Although body tracking is not available, this event coincided with a reversal in the animal's physical movement direction, from moving toward the lever from the central nosepoke before the lever press, to moving back to the central reward port after the lever press. In contrast, there is no reversal in the value modulation at the time of the lever press. The fact that the movement-related modulation (and not value modulation) followed the mice's movement direction during the trial further indicates that movement direction explains the choice selectivity in these DA neurons, and resists explanation in terms of RPE-related signaling.

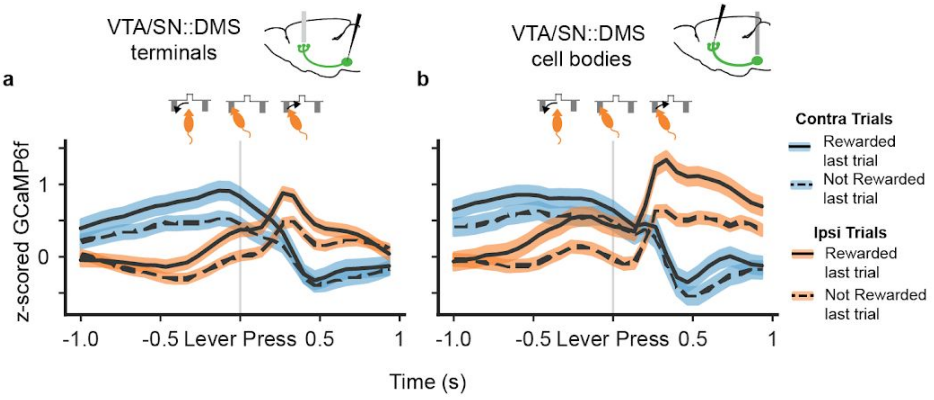


Figure 4: DA neurons that project to DMS reverse their choice selectivity after the lever press, around the time the mice reverse their movement direction. **(a)** GCaMP6f signal from VTA/SN::DMS terminals time-locked to the lever press, for contralateral choice trials (blue) and ipsilateral choice trials (orange), as well as rewarded (solid) and non-rewarded previous trial (dotted). The GCaMP6f traces for each choice crosses shortly after the lever-press, corresponding to the change in the mice's head direction around the time of the lever press (shown schematically above the graph). Colored fringes represent 1 standard error from activity averaged across recording sites (n = 12). **(b)** Same as **(a),** except VTA/SN::DMS cell body averaged across recording sites (n = 7) instead of terminals.

Discussion

Recent reports of qualitatively distinct DA responses, movement and RPE-related, have revived perennial puzzles about how the system contributes to both movement and reward, and more specifically raise the question whether there might be a unified computational description of both components in the spirit of the classic RPE models (Parker et al. 2016; Berke 2018; Coddington and Dudman 2018). To investigate these questions, we dissected movement and value selectivity in the responses of terminals and cell bodies of DMS-projecting DA neurons (**Figure 3**). Contrary to the hypothesis that movement related activity might reflect a RPE for contralateral value, multiple lines of evidence clearly indicated that the neurons instead contain distinct movement- and value-related signals, tied to different frames of reference. We did observe value-related signals preceding and following the leverpress, which were not previously appreciated in the DMS signal and which are consistent with the anticipatory component of a classic RPE response. But because these were oriented with respect to the value of the chosen action, not the contralateral one, they cannot explain the side-specific movement selectivity. The

two responses also showed clearly distinct time courses; in particular, the side selectivity reversed polarity following the leverpress, but value modulation did not.

The RPE account of the DA response has long held out hope for a unifying perspective on the system's dual roles in movement and reward, by proposing that the system's reward-related responses ultimately affect movement indirectly, either by driving learning about movement preferences (Montague, Dayan, and Sejnowski 1996) or by modulating motivation to act (Niv et al. 2007). It also accounts for multiple seemingly distinct components of the classic DA response, including anticipatory and reward-related responses, and responses to novel neutral cues. However, the present analyses clearly show that side-specific responses in DMS resist explanation in terms of an extended RPE account, and instead simply reflect planned or ongoing movements. Such a signal is, of course, well situated to play a direct role eliciting or executing contralateral movements, via differentially modulating the direct and indirect pathways out of the striatum (Alexander and Crutcher 1990; Collins and Frank 2014).

However, it is less clear how this signal might interact with the plasticity mechanisms hypothesized to be modulated by the RPE response (Reynolds and Wickens 2002); (Frank, Seeberger, and O'reilly 2004; Steinberg et al. 2013). For instance, how would recipient synapses distinguish an RPE component of the signal (appropriate for surprise-modulated learning) from a component more relevant to movement elicitation (Berke 2018)? One possibility is that plasticity in the dorsal striatum itself follows different rules, which might require an action rather than a prediction error signal; for instance, it has been suggested that some types of instrumental learning are correlational rather than error-driven (Doeller, King, and Burgess 2008) and in particular that habit learning in adjacent dorsolateral striatum might be based on directly recording elicited movements rather than error-driven learning of values (Miller, Shenhav, and Ludvig 2018). Overall, our results point to the need for an extended computational account that incorporates the movement signals as well as the RPE ones.

Another striking aspect of the results was the co-occurrence of two distinct frames of reference in the signal. Movement selectivity tracked choices contralateral versus ipsilateral of the recorded hemisphere – appropriate for motor control– but the value component instead related to the reward expected for the chosen, versus unchosen, action. This is suitable for a classic RPE for learning "state" values (since overall value expectancy at any point in time is conditioned on the choices the animal has made; (Morris et al. 2006), and also consistent with the bulk of BOLD responses in human neuroimaging, where value-related responding throughout dopaminergic targets tends to be organized on chosen-vs-unchosen lines (Daw et al. 2006; Boorman et al. 2009; Li and Daw 2011; O'Doherty 2014). At the same time, there have been persistent suggestions that given the high dimensionality of an organism's action space, distinct action-specific error signals would be useful for learning about different actions (Russell and Zimdars 2003; Frank and Badre 2012; Diuk et al. 2013) or types of predictions (Gershman and Schoenbaum 2017; Lau, Monteiro, and Paton 2017). Along these lines, there is evidence from BOLD neuroimaging for contralateral error and value signals in the human brain (Gershman, Pesaran, and Daw 2009; Palminteri et al. 2009). Though the current study finds no

evidence for such laterally decomposed RPEs in DMS, the decomposition of error signals remains an important possibility for future work aimed at understanding heterogeneity of dopamine responses, including also other anomalous response features like ramps (Mark W. Howe et al. 2013; Berke 2018; Gershman 2014; Hamid et al. 2016).

Interestingly, our results were consistent across at both recording sites with DMS-projecting DA neurons: the cell bodies and the terminals (**Figure 3d-f, Figure 4b**). This indicates that the movement selectivity was not introduced in DA neurons at the terminal level, e.g. via striatal cholinergic interneurons or glutamatergic inputs (Kosillo et al. 2016).

An important limitation of the study is the use of fiber photometry, which assesses bulk GCaMP6f responses at the recording site rather than resolving individual neurons. Thus it remains possible that individual neurons do not multiplex the two signals we observe, and that they are instead segregated between distinct populations. Future work should use higher resolution methods to examine these questions at the level of individual DA neurons. A related limitation of this study is the relatively coarse behavioral monitoring; notably, we infer that the reversal in selectivity seen in **Figure 4** reflects a change in movement direction, but head tracking would be required to verify this more directly. More generally, future work with finer instrumentation could usefully dissect response components related to finer-grained movements, and examine how these are related to (or dissociated from) value signals.

Methods

This article reports new analysis on data originally reported by (Parker et al. 2016). We briefly summarize the methods from that study here.

Mice and Surgeries

This article reports on data from 17 male TH::IRES-Cre mice, from which GCaMP6f recordings were obtained from DA neurons via fiber photometry. In the case of VTA/SN::DMS cell body recordings, Cre-dependent GCaMP6f virus was injected into the DMS, and fibers were placed on the cell bodies in VTA/SNc, enabling recordings from retrogradely labeled cells (n=4 mice). In the case of DA terminal recordings, Cre-dependent GCaMP6f virus was injected into the VTA/SNc, and fibers were placed in the DMS (n=12 mice). The recording hemisphere was counterbalanced across mice.

Instrumental Reversal Learning Task

The recordings were obtained while the mice performed a reversal learning task in an operant chamber with a central nose poke, retractable levers on each side of the nose poke, and reward delivery in a receptacle beneath the central nose poke.

Each trial began with the illumination of the center nose port. After the mice entered the nose port, the two levers were presented with a delay that varied between 0-1 seconds. The mice then had 10 seconds to press a lever, otherwise the trial was classified as an abandoned trial and excluded from analysis (this amounted to <2 % of trials for all mice). After the lever-press, an additional random 0-1 second delay (0.1 second intervals, uniform distribution) preceded either CS- with no reward delivery or CS+ with a 4µl reward of 10% sucrose in H20. Reward outcomes were accompanied by different auditory stimulus: 0.5 seconds of white noise for CS- and 0.5 seconds of 5 kHz pure tone for CS+. Every trial ended with a constant 3 seconds inter-trial delay.

Data Post-processing

All fiber photometry recordings were acquired at 15 Hz. 1-5 recording sessions were obtained per recording site (1 session/day), and these recordings were concatenated across session for all analyses. The signal from each recording site were post-processed with a high-pass FIR filter with a passband of 0.4 Hz, stopband of 0.1 Hz, and a stopband attenuation of 10 dB to remove baseline fluorescence and correct drift in baseline. We derived dF/F by dividing the high-pass filtered signal by the mean of the signal before high-pass filtering. We then z-scored dF/F for each recording site, with the the mean and standard error calculated for the entire recording from each site.

The VTA/SN::DMS terminals data consisted of (10108 total trials across 12 recording sites) and VTA/SN::DMS cell-bodies (4938 total trials across 7 recording sites).

Q Learning Mixed Effect Model

We fit a trial-by-trial Q-learning mixed effect model to the behavioral data from each of the 12 mice on all recording sites, and combined data across mice with a hierarchical model. The model was initialized with a Q value of 0 for each action and updated at each trial according to:

$$Q_{t+1}(c_t) = Q_t(c_t) + \alpha(r_t - Q_t(c_t))$$

where Q is the value for both options, c_t is the option chosen on trial *t* (lever either contralateral or ipsilateral to recording site), and 0 <= α <= 1 is a free learning rate parameter. The subject's probability to choose choice *c* was then given by a softmax equation:

$$P(c_t = c) \propto exp(\beta \cdot Q_t(c) + stay \cdot I(c, c_{t-1}))$$

where β is a free inverse temperature parameter, *stay* is a free parameter encoding how likely the animal will repeat its choice from the last trial, and *I* is a binary indicator function for choice repetition (1 if *c* was chosen on the previous trial; 0 otherwise). The three free parameters of the

model were estimated separately for each subject, but jointly (in a hierarchical random effects model) with group-level mean and variance parameters reflecting the distribution, over the population, of each subject-level parameter.

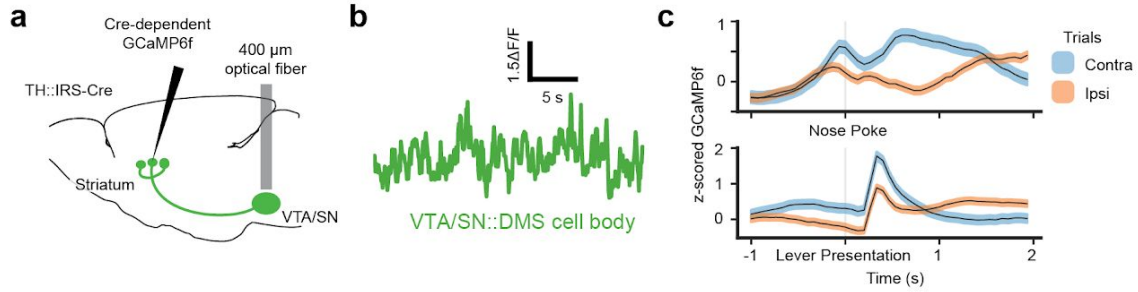
The parameters were estimated using Hamiltonian Monte Carlo, as implemented in the Stan programming language (version 2.17.1.0; (Carpenter et al. 2017)). Samples from the posterior distribution over the parameters were extracted using the Python package PyStan (Carpenter et al. 2017). We ran the model with 4 chains of 1,000 iterations for each (of which the first 250 were discarded for burn-in), and the parameter adapt_delta set to 0.99. We verified convergence by visual inspection and by verifying that the potential scale reduction statistic Rhat (Gelman and Rubin 1992) was close to 1.0 (<0.003 for all parameters).

We used the sampled parameters to compute per-trial Q values for each action, trial, and mouse. We calculated the difference between the Q values of the chosen action and unchosen action for each trial. We binned the difference of these Q values for each trial and plotted the average GCaMP6f time-locked to lever presentation for each bin (**Figure 2b**).

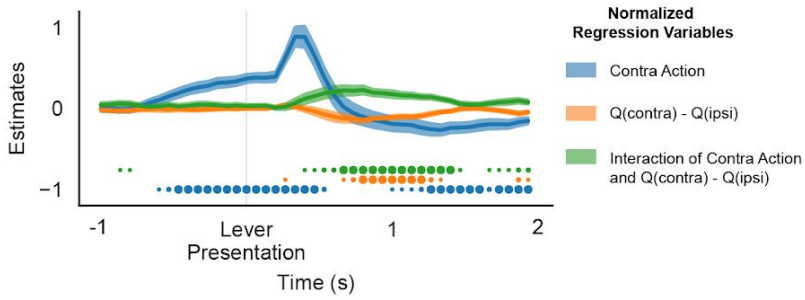
Regression Model

In **Figure 3c,f**, we performed a linear mixed effect model regression to predict GCaMP6f signal at each time point based on Q-values, choice (contralateral vs ipsilateral), their interaction, and an intercept. We took the difference of Q values for the chosen vs unchosen levers, then we standardized the difference of Q values for each mice and each recording site. GCaMP6f was time-locked to lever presentation, regressing to data points 1 second before and 2 seconds after the time-locked event for 45 total regressions. The regression, as well as the calculation of p values, was performed with the MixedModels package in Julia (Bezanson et al. 2014).

Supplementary Figures



Supplement Figure 1: Recording from VTA/SN::DMS cell bodies (n = 7) (**a**) Surgical schematic for recording with optical fibers from the GCaMP6f VTA/SN::DMS cell-bodies. Projections were determined using viral traces. (**b**) Sample GCaMP6f traces from VTA/SN::DMS cell bodies. **(c, d)** We also see contralateral choice selectivity in DMS DA cell bodies when the signals are time-locked to nose poke (**c**) and lever presentation (**d**).



Supplement Figure 2: Mixed effect model regression on each datapoint from 3 seconds of GCaMP6f traces of VTA/SN::DMS terminals (n = 12). Explanatory variables include the action of the mice (blue), the difference in Q values for contralateral and ipsilateral choices (orange), their interaction (green), and an intercept. Colored fringes represent 1 standard error from estimates. Dots at bottom mark timepoints where the corresponding effect is significantly different from zero at p<.05 (small dot), p<.01 (medium dot), p<.001(large dot).

References List

Alexander, G. E., and M. D. Crutcher. 1990. “Functional Architecture of Basal Ganglia Circuits: Neural Substrates of Parallel Processing.” *Trends in Neurosciences* 13 (7): 266–71.

Baird, L. C. 1994. “Reinforcement Learning in Continuous Time: Advantage Updating.” In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*, 4:2448–53 vol.4.

Barter, Joseph W., Suellen Li, Dongye Lu, Ryan A. Bartholomew, Mark A. Rossi, Charles T. Shoemaker, Daniel Salas-Meza, Erin Gaidis, and Henry H. Yin. 2015. “Beyond Reward Prediction Errors: The Role of Dopamine in Movement Kinematics.” *Frontiers in Integrative Neuroscience* 9 (May): 39.

Barto, A. G., R. S. Sutton, and C. W. Anderson. 1983. “Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems.” *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13 (5): 834–46.

Barto, Andrew G. 1995. “1‘ 1 Adaptive Critics and the Basal Ganglia,.’” *Models of Information Processing in the Basal Ganglia*, 215.

Berke, Joshua D. 2018. “What Does Dopamine Mean?” *Nature Neuroscience* 21 (6): 787–93.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2014. “Julia: A Fresh Approach to Numerical Computing.” *arXiv [cs.MS]*. arXiv. http://arxiv.org/abs/1411.1607.

Boorman, Erie D., Timothy E. J. Behrens, Mark W. Woolrich, and Matthew F. S. Rushworth. 2009. “How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action.” *Neuron* 62 (5): 733–43.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. “Stan: A Probabilistic Programming Language.” *Journal of Statistical Software, Articles* 76 (1): 1–32.

Coddington, Luke T., and Joshua T. Dudman. 2018. “The Timing of Action Determines Reward Prediction Signals in Identified Midbrain Dopamine Neurons.” *Nature Neuroscience*, October. https://doi.org/10.1038/s41593-018-0245-7.

Cohen, Jeremiah Y., Sebastian Haesler, Linh Vong, Bradford B. Lowell, and Naoshige Uchida. 2012. “Neuron-Type-Specific Signals for Reward and Punishment in the Ventral Tegmental Area.” *Nature* 482 (7383): 85–88.

Collins, Anne G. E., and Michael J. Frank. 2014. “Opponent Actor Learning (OpAL): Modeling Interactive Effects of Striatal Dopamine on Reinforcement Learning and Choice Incentive.” *Psychological Review* 121 (3): 337–66.

Daw, Nathaniel D., John P. O’Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. 2006. “Cortical Substrates for Exploratory Decisions in Humans.” *Nature* 441 (7095): 876–79.

Diuk, Carlos, Karin Tsai, Jonathan Wallis, Matthew Botvinick, and Yael Niv. 2013. “Hierarchical Learning Induces Two Simultaneous, but Separable, Prediction Errors in Human Basal Ganglia.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 33 (13): 5797–5805.

Dodson, Paul D., Jakob K. Dreyer, Katie A. Jennings, Emilie C. J. Syed, Richard Wade-Martins, Stephanie J. Cragg, J. Paul Bolam, and Peter J. Magill. 2016. “Representation of Spontaneous Movement by Dopaminergic Neurons Is Cell-Type Selective and Disrupted in Parkinsonism.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (15): E2180–88.

Doeller, Christian F., John A. King, and Neil Burgess. 2008. “Parallel Striatal and Hippocampal

Systems for Landmarks and Boundaries in Spatial Memory." *Proceedings of the National Academy of Sciences of the United States of America* 105 (15): 5915–20.

Frank, Michael J., and David Badre. 2012. "Mechanisms of Hierarchical Reinforcement Learning in Corticostriatal Circuits 1: Computational Analysis." *Cerebral Cortex* 22 (3): 509–26.

Frank, Michael J., Lauren C. Seeberger, and Randall C. O'reilly. 2004. "By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism." *Science* 306 (5703): 1940–43.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 7 (4): 457–72.

Gershman, Samuel J. 2014. "Dopamine Ramps Are a Consequence of Reward Prediction Errors." *Neural Computation* 26 (3): 467–71.

Gershman, Samuel J., Bijan Pesaran, and Nathaniel D. Daw. 2009. "Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (43): 13524–31.

Gershman, Samuel J., and Geoffrey Schoenbaum. 2017. "Rethinking Dopamine Prediction Errors." *bioRxiv*. https://doi.org/10.1101/239731.

Hamid, Arif A., Jeffrey R. Pettibone, Omar S. Mabrouk, Vaughn L. Hetrick, Robert Schmidt, Caitlin M. Vander Weele, Robert T. Kennedy, Brandon J. Aragona, and Joshua D. Berke. 2016. "Mesolimbic Dopamine Signals the Value of Work." *Nature Neuroscience* 19 (1): 117–26.

Howe, Mark W., Patrick L. Tierney, Stefan G. Sandberg, Paul E. M. Phillips, and Ann M. Graybiel. 2013. "Prolonged Dopamine Signalling in Striatum Signals Proximity and Value of Distant Rewards." *Nature* 500 (7464): 575–79.

Howe, M. W., and D. A. Dombeck. 2016. "Rapid Signalling in Distinct Dopaminergic Axons during Locomotion and Reward." *Nature* 535 (7613): 505–10.

Kosillo, Polina, Yan-Feng Zhang, Sarah Threlfell, and Stephanie J. Cragg. 2016. "Cortical Control of Striatal Dopamine Transmission via Striatal Cholinergic Interneurons." *Cerebral Cortex* , August. https://doi.org/10.1093/cercor/bhw252.

Lau, Brian, Tiago Monteiro, and Joseph J. Paton. 2017. "The Many Worlds Hypothesis of Dopamine Prediction Error: Implications of a Parallel Circuit Architecture in the Basal Ganglia." *Current Opinion in Neurobiology* 46 (October): 241–47.

Li, Jian, and Nathaniel D. Daw. 2011. "Signals in Human Striatum Are Appropriate for Policy Update rather than Value Prediction." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 31 (14): 5504–11.

Miller, Kevin, Amitai Shenhav, and Elliot Ludvig. (2018). "Habits without Values." *Psychological Review*. Accessed September 18, 2018. https://doi.org/10.1101/067603.

Montague, P. R., P. Dayan, and T. J. Sejnowski. 1996. "A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 16 (5): 1936–47.

Morris, Genela, Alon Nevet, David Arkadir, Eilon Vaadia, and Hagai Bergman. 2006. "Midbrain Dopamine Neurons Encode Decisions for Future Action." *Nature Neuroscience* 9 (8): 1057–63.

Niv, Yael, Nathaniel D. Daw, Daphna Joel, and Peter Dayan. 2007. "Tonic Dopamine: Opportunity Costs and the Control of Response Vigor." *Psychopharmacology* 191 (3): 507–20.

O'Doherty, John P. 2014. "The Problem with Value." *Neuroscience and Biobehavioral Reviews* 43 (June): 259–68.

O'Doherty, John P., Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J. Dolan. 2004. "Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning." *Science* 304 (5669): 452–54.

Palminteri, Stefano, Thomas Boraud, Gilles Lafargue, Bruno Dubois, and Mathias Pessiglione. 2009. "Brain Hemispheres Selectively Track the Expected Value of Contralateral Options." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 29 (43): 13465–72.

Parker, Nathan F., Courtney M. Cameron, Joshua P. Taliaferro, Junuk Lee, Jung Yoon Choi, Thomas J. Davidson, Nathaniel D. Daw, and Ilana B. Witten. 2016. "Reward and Choice Encoding in Terminals of Midbrain Dopamine Neurons Depends on Striatal Target." *Nature Neuroscience* 19 (6): 845–54.

Reynolds, John N. J., and Jeffery R. Wickens. 2002. "Dopamine-Dependent Plasticity of Corticostriatal Synapses." *Neural Networks: The Official Journal of the International Neural Network Society* 15 (4-6): 507–21.

Roesch, Matthew R., Donna J. Calu, and Geoffrey Schoenbaum. 2007. "Dopamine Neurons Encode the Better Option in Rats Deciding between Differently Delayed or Sized Rewards." *Nature Neuroscience* 10 (12): 1615–24.

Russell, Stuart, and Andrew L. Zimdars. 2003. "Q-Decomposition for Reinforcement Learning Agents." In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 656–63. ICML'03. Washington, DC, USA: AAAI Press.

Samuelson, P. A. 1938. "A Note on the Pure Theory of Consumer's Behaviour." *Economica* 5 (17): 61–71.

Schultz, W., P. Dayan, and P. R. Montague. 1997. "A Neural Substrate of Prediction and Reward." *Science* 275 (5306): 1593–99.

Silva, Joaquim Alves da, Fatuel Tecuapetla, Vitor Paixão, and Rui M. Costa. 2018. "Dopamine Neuron Activity before Action Initiation Gates and Invigorates Future Movements." *Nature* 554 (7691): 244–48.

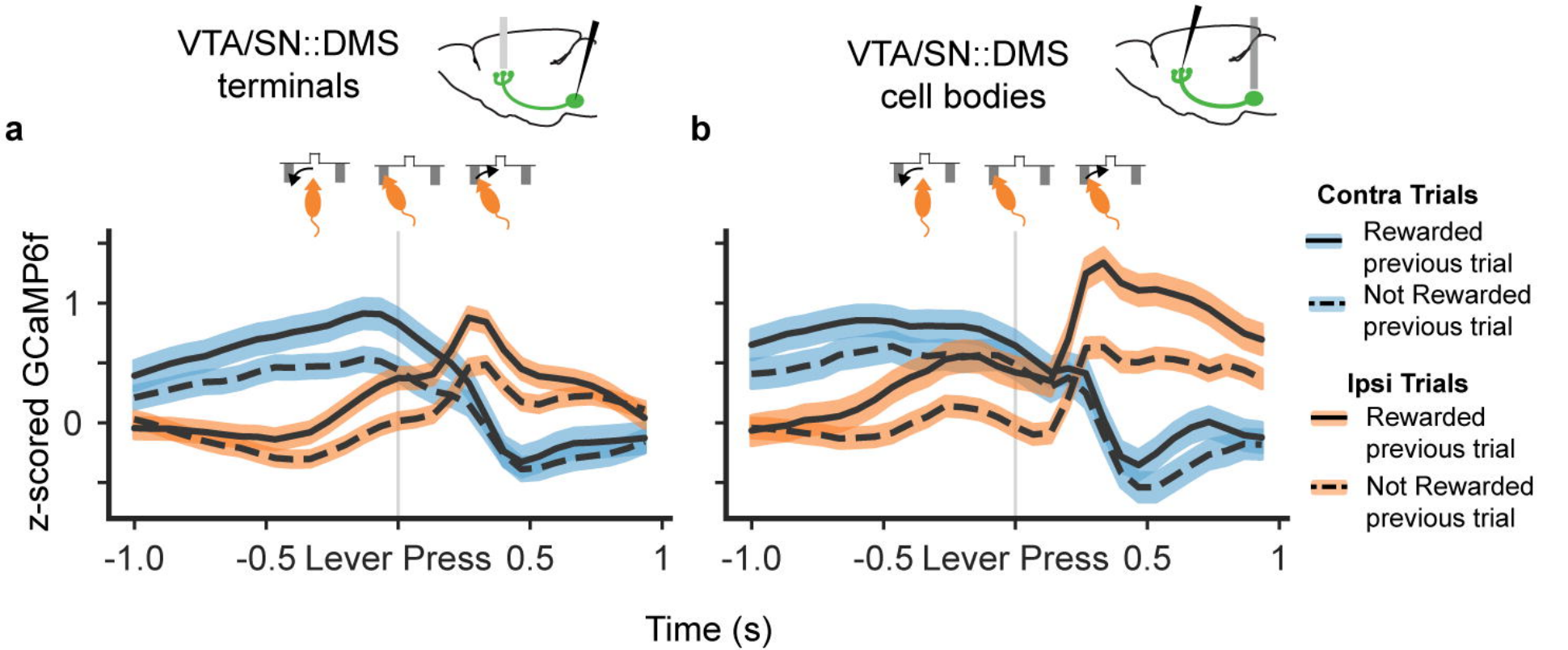
Soares, Sofia, Bassam V. Atallah, and Joseph J. Paton. 2016. "Midbrain Dopamine Neurons Control Judgment of Time." *Science* 354 (6317): 1273–77.

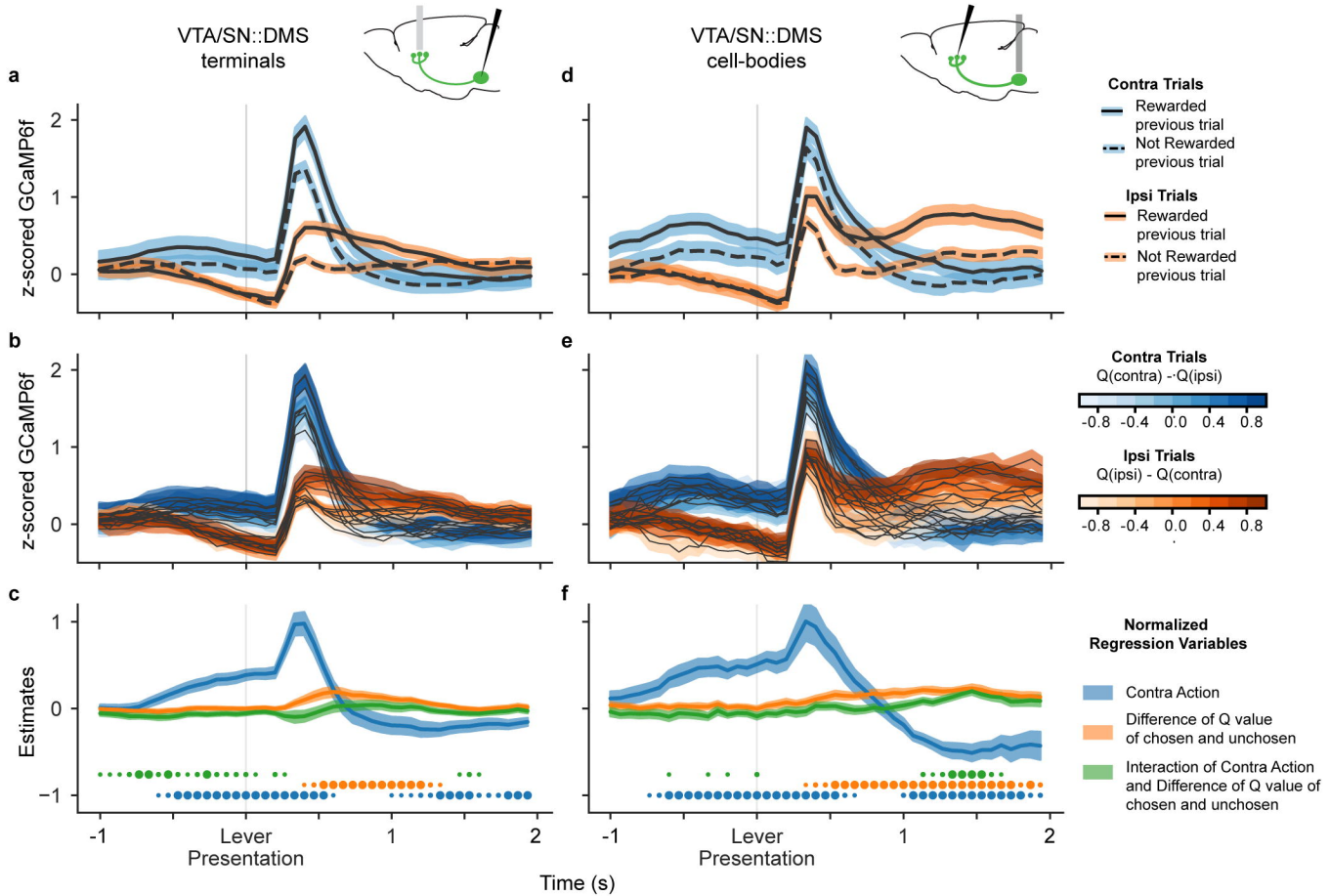
Stan Development Team. 2017. *StataStan: the Stata interface to Stan*. http://mc-stan.org

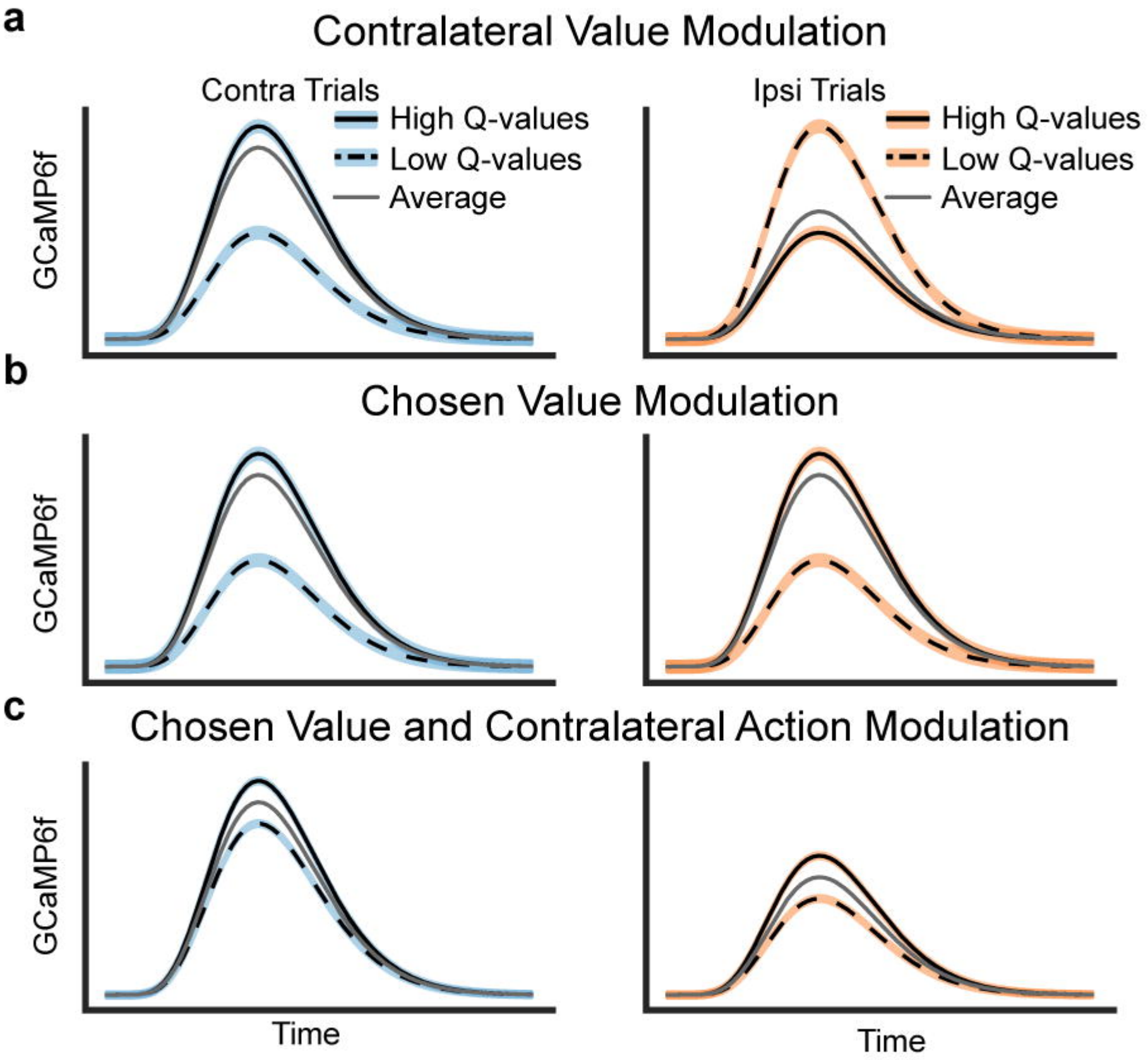
Steinberg, Elizabeth E., Ronald Keiflin, Josiah R. Boivin, Ilana B. Witten, Karl Deisseroth, and Patricia H. Janak. 2013. "A Causal Link between Prediction Errors, Dopamine Neurons and Learning." *Nature Neuroscience* 16 (7): 966–73.

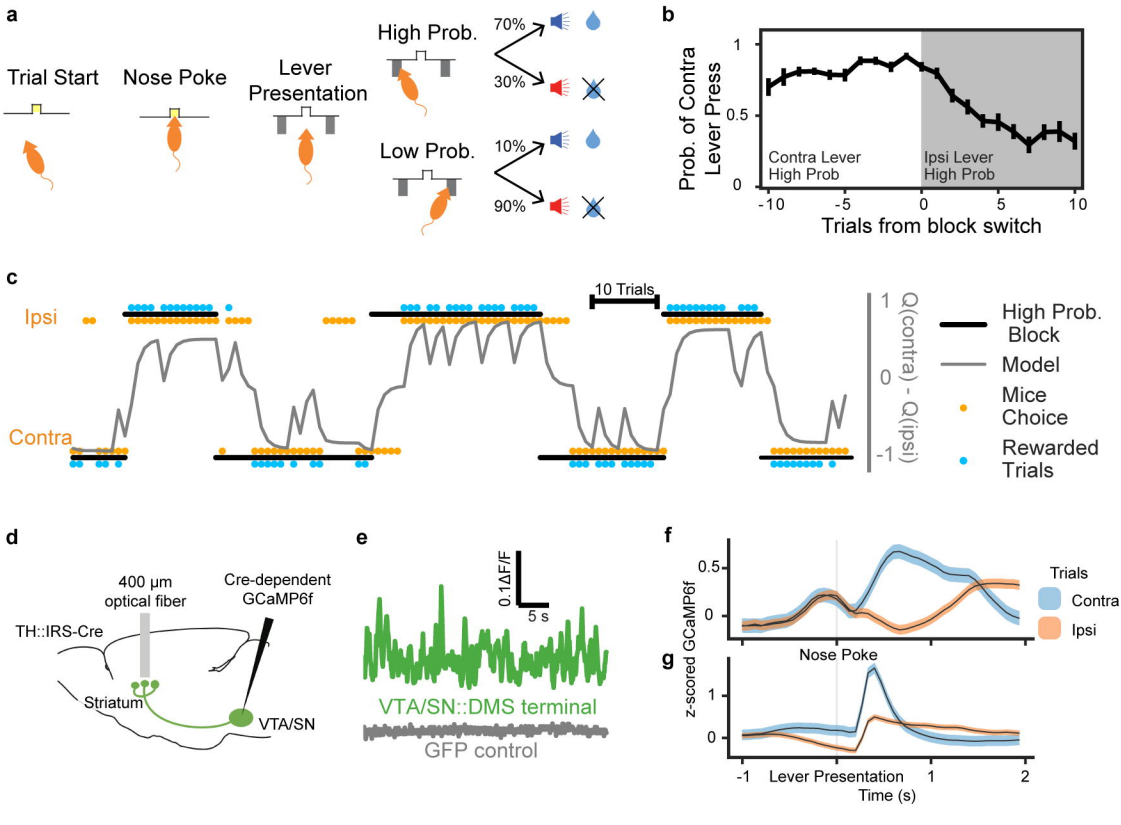
Takahashi, Yuji, Geoffrey Schoenbaum, and Yael Niv. 2008. "Silencing the Critics: Understanding the Effects of Cocaine Sensitization on Dorsolateral and Ventral Striatum in the Context of an Actor/critic Model." *Frontiers in Neuroscience* 2 (1): 86–99.

Wise, Roy A. 2004. "Dopamine, Learning and Motivation." *Nature Reviews. Neuroscience* 5 (6): 483–94.


VTA/SN::DMS
terminals
VTA/SN::DMS
cell bodies
a
b
z-scored GCaMP6f
1
0
-1.0
-0.5
Lever Press
0.5
1
Time (s)
Contra Trials
Rewarded
previous trial
Not Rewarded
previous trial
Ipsi Trials
Rewarded
previous trial
Not Rewarded
previous trial


VTA/SN::DMS
terminals
VTA/SN::DMS
cell-bodies
a
b
c
d
e
f
z-scored GCaMP6f
Estimates
Contra Trials
Rewarded
previous trial
Not Rewarded
previous trial
Ipsi Trials
Rewarded
previous trial
Not Rewarded
previous trial
Contra Trials
Q(contra) - Q(ipsi)
-0.8 -0.4 0.0 0.4 0.8
Ipsi Trials
Q(ipsi) - Q(contra)
-0.8 -0.4 0.0 0.4 0.8
Normalized
Regression Variables
Contra Action
Difference of Q value
of chosen and unchosen
Interaction of Contra Action
and Difference of Q value of
chosen and unchosen
-1
Lever
Presentation
1
2
Time (s)


a
Contralateral Value Modulation
Contra Trials
Ipsi Trials
High Q-values
Low Q-values
Average
GCaMP6f
b
Chosen Value Modulation
c
Chosen Value and Contralateral Action Modulation
Time


a
Trial Start
Nose Poke
Lever Presentation
High Prob.
70%
30%
Low Prob.
10%
90%
b
Prob. of Contra Lever Press
1
0.5
0
Contra Lever High Prob
Ipsi Lever High Prob
-10
-5
0
5
10
Trials from block switch
c
Ipsi
Contra
10 Trials
1
0
-1
Q(contra) - Q(ipsi)
High Prob. Block
Model
Mice Choice
Rewarded Trials
d
400 µm optical fiber
Cre-dependent GCaMP6f
TH::IRS-Cre
Striatum
VTA/SN
e
0.1ΔF/F
5 s
VTA/SN::DMS terminal
GFP control
f
0.5
0
Trials
Contra
Ipsi
g
z-scored GCaMP6f
Nose Poke
1
0
-1
Lever Presentation
1
2
Time (s)