Striatal action-value neurons reconsidered

Lotem Elber-Dorozko\textsuperscript{1} and Yonatan Loewenstein\textsuperscript{1,2}

\textsuperscript{1}The Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem.

\textsuperscript{2}Department of Neurobiology, The Alexander Silberman Institute of Life Sciences and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem.

Corresponding author and lead contact: Lotem.elber@mail.huji.ac.il

Abstract

It is generally believed that during economic decisions, striatal neurons represent the values associated with different actions. This hypothesis is based on studies, in which the activity of striatal neurons was measured while the subject was learning to prefer the more rewarding action. Here we show that these publications are subject to at least one of two critical confounds. First, we show that even weak temporal correlations in the neuronal data may result in an erroneous identification of action-value representations. Second, we show that experiments and analyses designed to dissociate action-value representation from the representation of other decision variables cannot do so. We suggest solutions to identifying action-value representation that are not subject to these confounds. Applying one solution to previously identified action-value neurons in the basal ganglia we fail to detect action-value representations. We conclude that the claim that striatal neurons encode action-values must await new experiments and analyses.

Key words

action-value; striatum; temporal correlations; statistical confounds
There is a long history of operant learning experiments, in which a subject, human or animal, repeatedly chooses between actions and is rewarded according to its choices. A popular theory posits that the subject’s decisions in these tasks utilize estimates of the different *action-values*. These action-values correspond to the expected reward associated with each of the actions, and actions associated with a higher estimated action-value are more likely to be chosen (Sutton & Barto, 1998). In recent years, there is a lot of interest in the neural mechanisms underlying this computation (Louie & Glimcher, 2012; Schultz, 2015). In particular, based on electrophysiological, functional magnetic resonance imaging (fMRI) and intervention experiments, it is now widely accepted that a population of neurons in the striatum represents these action-values, adding sway to this action-value theory (Cai, Kim, & Lee, 2011; Fitzgerald, Friston, & Dolan, 2012; Funamizu, Ito, Doya, Kanzaki, & Takahashi, 2015; Guitart-Masip et al., 2012; Her, Huh, Kim, & Jung, 2016; Ito & Doya, 2009, 2015a, 2015b; H. Kim, Lee, & Jung, 2013; H. Kim, Sul, Huh, Lee, & Jung, 2009; S. Kim, Cai, Hwang, & Lee, 2012; Y. B. Kim et al., 2007; Lau & Glimcher, 2008; Lee, Seo, Monte, & Averbeck, 2015; Samejima, Ueda, Doya, & Kimura, 2005; Stalnaker, Calhoon, Ogawa, Roesch, & Schoenbaum, 2010; Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012; Wang, Miura, & Uchida, 2013; Wunderlich, Rangel, & O’Doherty, 2009). Here we challenge the evidence for action-value representation in the striatum by describing two major confounds that have been successfully addressed when analyzing the data.

To identify neurons that represent the internal values of the different actions, researchers have searched for neurons whose firing rate is significantly correlated with the average reward associated with exactly one of the actions. There are several ways of defining the average reward associated with an action. For example, the average reward can be defined by the reward schedule, e.g., the probability of a reward associated with the action. Alternatively, one can adopt
the subject’s perspective, and use the subject-specific history of rewards and actions in order to estimate the average reward. In particular, the Rescorla–Wagner model (equivalent to the standard ones-state Q-learning model) has been used to estimate action-values (H. Kim et al., 2009; Samejima et al., 2005). In this model, the value associated with an action $i$ in trial $t$, termed $Q_i(t)$, is an exponentially-weighted average of the rewards associated with this action in past trials:

$$Q_i(t + 1) = Q_i(t) + \alpha (R(t) - Q_i(t)) \quad \text{if} \ a(t) = i$$

$$Q_i(t + 1) = Q_i(t) \quad \text{if} \ a(t) \neq i$$

where $a(t)$ and $R(t)$ denote the choice and reward in trial $t$, respectively, and $\alpha$ is the learning rate.

In a two-alternative task, the probability of choosing an action is a sigmoidal function, typically softmax, of the difference of the action-values (see also (Shteingart & Loewenstein, 2014)):

$$\Pr(a(t) = 1) = \frac{1}{1 + e^{-\beta(Q_1(t) - Q_2(t))}}$$

where $\beta$ is a parameter that determines the bias towards the action associated with the higher action-value. The parameters of the model, $\alpha$ and $\beta$, can be estimated from the behavior, allowing the researchers to compute $Q_1$ and $Q_2$ on a trial-by-trial basis.

In principle, one can identify the neurons that represent an action-value by identifying neurons for which the regression of the trial-by-trial spike count on one of the variables $Q_i(t)$ is statistically significant. Using this framework, electrophysiological studies have found that the firing rate of a substantial fraction of striatal neurons (12%-40% for different significance thresholds) is significantly correlated with an action-value. These and similar results were
considered as evidence that neurons in the striatum represent action-values (Funamizu et al., 2015; Her et al., 2016; Ito & Doya, 2015a, 2015b, H. Kim et al., 2013, 2009; Lau & Glimcher, 2008; Samejima et al., 2005).

In this paper we conduct a systematic literature search and conclude that the literature has, by and large, ignored two major confounds in this analysis. First, it is well-known that spurious correlations can emerge in correlation analysis if the variables are temporally correlated (Granger & Newbold, 1974; Phillips, 1986). Here we show that neurons can be erroneously identified as representing action-values when their firing rates are weakly temporally correlated. Second, it is also well-known that lack of a statistical significance in the analysis does not imply lack of correlation. Because in standard analyses neurons are classified as representing action-values if they have a significant regression coefficient on exactly one action-value and because decision-related variables such as policy are correlated with both action-values, neurons representing other decision-related variables may be misclassified as representing action-values. We propose different approaches to address these issues. Applying one of them to recordings from the basal ganglia, we fail to identify any action-value representation there. Thus, we conclude that the hypothesis that striatal neurons represent action-values still remains to be tested by experimental designs and analyses that are not subject to these confounds. In the Discussion we address additional conceptual issues with identifying such a representation.

This paper discusses a methodological problem that may also be of relevance in other fields of biology in general and neuroscience in particular. Nevertheless, the focus of this paper is a single scientific claim, namely, that action-value representation in the striatum is an established fact. Our criticism is restricted to the representation of action-values, and we do not make any claims regarding the possible representations of other decision variables, such as policy, chosen-value or
reward-prediction-error. This we leave for future studies. Moreover, we do not make any claims about the possible representations of action-values elsewhere in the brain, although our results suggest caution when looking for such representations.

The paper is organized in the following way. We commence by describing a standard method for identifying action-value neurons. Next, we show that this method erroneously classifies simulated neurons whose activity is temporally correlated as representing action-values. We show that this confound brings into question the conclusion of many existing publications. Then, we propose different methods for identifying action-value neurons, that overcome this confound. Applying such a method to basal ganglia recordings, in which action-value neurons were previously identified, we fail to conclusively detect any action-value representations. We continue by discussing the second confound: neurons that encode the policy (the probability of choice) may be erroneously identified as representing action-value, even when the policy is the result of learning algorithms that are devoid of action-value calculation. Then we discuss a possible solution to this confound.

**Results**

**Identifying action-value neurons**

We commence by examining the standard methods for identifying action-value neurons using a simulation of an operant learning experiment. We simulated a task, in which the subject repeatedly chooses between two alternative actions, which yield a binary reward with a probability that depends on the action. Specifically, each session in the simulation was composed of four blocks such that the probabilities of rewards were fixed within a block and varied between the blocks. The probabilities of reward in the blocks were (0.1,0.5), (0.9,0.5), (0.5,0.9)
and (0.5,0.1) for actions 1 and 2, respectively (Fig. 1A). The order of blocks was random and a block terminated when the more rewarding action was chosen more than 14 times within 20 consecutive trials (Ito & Doya, 2015a; Samejima et al., 2005).

To simulate learning behavior, we used the Q-learning framework (Eqs. (1) and (2) with $\alpha = 0.1$ and $\beta = 2.5$ (taken from distributions reported in (H. Kim et al., 2009)) and initial conditions $Q_1(1) = 0.5$). As demonstrated in Fig. 1A, the model learned, such that the probability of choosing the more rewarding alternative increased over trials (black line). To model the action-value neurons, we simulated neurons whose firing rate is a linear function of one of the two Q-values and whose spike count in a 1 sec trial is randomly drawn from a corresponding Poisson distribution (see Materials and Methods). The firing rates and spike counts of two such neurons, representing action-values 1 and 2, are depicted in Fig. 1B in red and blue, respectively.

One standard method for identifying action-value neurons is to compare the spike counts after learning, at the end of the blocks (horizontal bars in Fig. 1B). Considering the red-labeled Poisson neuron, the spike count in the last 20 trials of the second block, in which the probability of reward associated with action 1 was 0.9, was significantly higher than that count in the first block, in which the probability of reward associated with action 1 was 0.1 ($p < 0.01$; rank sum test). By contrast, there was no significant difference in the spike counts between the third and fourth blocks, in which the probability of reward associated with action 1 was equal ($p = 0.91$; rank sum test; Fig. 1B, red). This is consistent with the fact that the red-labeled neuron was an action 1-value neuron: its firing rate was a linear function of the value of action 1. Similarly for the blue labeled neuron, the spike counts in the last 20 trials of the first two blocks were not significantly different ($p = 0.92$; rank sum test), but there was a significant difference in the counts between the third and fourth blocks ($p < 0.001$; rank sum test). These results are
consistent with the probabilities of reward associated with action 2 and the fact that in our simulations, this neuron’s firing rate was modulated by the value of action 2 (Fig. 1B, blue).

This approach for identifying action-value neurons is limited, however, for several reasons. First, it considers only a fraction of the data, the last 20 trials in a block. Second, action-value neurons are not expected to represent the block average probabilities of reward. Rather, they will represent a subjective estimate, which is based on the subject-specific history of actions and rewards. Therefore, it is more common to identify action-value neurons by regressing the spike count on subjective action-values, estimated from the subject’s history of choices and rewards (Funamizu et al., 2015; Ito & Doya, 2015a, 2015b; H. Kim et al., 2009; Lau & Glimcher, 2008; Samejima et al., 2005). Note that when studying behavior in experiments, we have no direct access to these estimated action-values, in particular because the values of the parameters α and β are unknown. Therefore, following common practice, we estimated the values of α and β from the model’s sequence of choices and rewards using maximum likelihood, and used the estimated learning rate (α) and the choices and rewards to estimate the action-values (thin lines in Fig. 1C, see Materials and Methods). These estimates were similar to the true action-value, which underlay the model’s choice behavior (thick lines in Fig. 1C).

Next, we regressed the spike count of each simulated neuron on the two estimated action-values from its corresponding session. As expected, the t-values of the regression coefficients of the red-labeled action 1-value neuron was significant for the estimated $Q_1$ ($t_{182}(Q_1) = 4.05$) but not for the estimated $Q_2$ ($t_{182}(Q_2) = -0.27$). Similarly, the t-values of the regression coefficients of the blue-labeled action 2-value neuron was significant for the estimated $Q_2$ ($t_{182}(Q_2) = 3.05$) but not for the estimated $Q_1$ ($t_{182}(Q_1) = 0.78$).
A population analysis of the t-values of the two regression coefficients is depicted in Fig. 1D,E. As expected, a substantial fraction (42%) of the simulated neurons were identified as action-value neurons. Only 2% of the simulated neurons had significant regression coefficients with both action-values. Such neurons are typically classified as state ($\Sigma Q$) or policy (also known as preference) ($\Delta Q$) neurons, if the two regression coefficients have the same or different signs, respectively (Ito & Doya, 2015a). Note that despite the fact that by construction, all neurons were action-value neurons, not all of them were detected as such by this method. This failure occurred for two reasons. First, the estimated action-values are not identical to the true action-values, which determine the firing rates. This is because of the finite number of trials and the stochasticity of choice (note the difference, albeit small, between the thin and thick lines in Fig. 1C). Second and more importantly, the spike count in a trial is only a noisy estimate of the firing rate because of the Poisson generation of spikes.
**Figure 1** Model of action-value neurons (A) Behavior of the model in an example session, composed of four blocks (separated by dashed vertical lines). The probabilities of reward for choosing actions 1 and 2 are denoted by the pair of numbers above the block. Black line denotes the probability of choosing action 1; vertical lines denote choices in individual trials, where red and blue denote actions 1 and 2, respectively, and long and short lines denote rewarded and unrewarded trials, respectively. (B) Neural activity. Firing rate (line) and spike-count (dots) of two example simulated action-value neurons in the session depicted in (A). The red and blue-labeled neurons represent $Q_1$ and $Q_2$, respectively. Black horizontal lines denote the mean spike count in the last 20 trials of the block. Error bars denote the standard error of the mean. The two asterisks denote $p<0.01$ (rank sum test). (C) Values. Thick red and blue lines denote $Q_1$ and $Q_2$, respectively. Note that the firing rates of the two neurons in (B) are a linear function of these values. Thin red and blue lines denote the estimates of $Q_1$ and $Q_2$, respectively. The similarity between the thick and thin lines indicates that the parameters of the model can be accurately estimated from the behavior (see also Materials and Methods). (D) and (E) Population analysis. (D) Example of 500 simulated action-value neurons from randomly chosen sessions. Each dot corresponds to a single neuron and the coordinates correspond to the t-values of regression of the spike counts on the estimated values of the two actions. Dashed lines at t=2 denote the significance boundaries. Color of dots denote significance: dark red and blue denote a significant regression coefficient only on one estimated action-value, action 1 or action 2, respectively; light blue – significant regression coefficients on both estimated action-values with similar signs ($\Sigma Q$), orange - significant regression coefficients on both estimated action-values with opposite signs ($\Delta Q$). Black – no significant regression coefficients. The two simulated neurons in (B) are denoted by squares. (E) Fraction of neurons in each category, estimated from 20,000 simulated neurons in 1,000 sessions. Error bars denote the standard error of the mean.
Dashed lines denote the naïve expected false positive rate from the significance threshold (see Materials and Methods).
Several prominent studies have implemented the methods we described in this section and reported that a substantial fraction (10-40% depending on significance threshold) of striatal neurons represent action-values (Ito & Doya, 2015a, 2015b; Samejima et al., 2005). In the next two sections we show that these methods, and similar methods employed by other studies (Cai et al., 2011; Fitzgerald et al., 2012; Funamizu et al., 2015; Guitart-Masip et al., 2012; Her et al., 2016; Ito & Doya, 2009; H. Kim et al., 2013, 2009; S. Kim et al., 2012; Y. B. Kim et al., 2007; Lau & Glimcher, 2008; Stalnaker et al., 2010; Wang et al., 2013; Wunderlich et al., 2009) are all subject to at least one of two major confounds.

**Confound 1 – Temporal correlations**

Simulated random-walk neurons are erroneously identified as action-value neurons

The red and blue-labeled neurons in Fig. 1D were classified as action-value neurons because their t-values were improbable under the null hypothesis that the firing rate of the neuron is not modulated by action-values. The significance threshold ($t=2$) was computed assuming that trials are independent in time. To see why this assumption is essential, we consider a case in which it is violated. Fig. 2A depicts the firing rates and spike counts of two simulated Poisson neurons, whose firing rates follow a bounded Gaussian random-walk process:

$$f(t + 1) = [f(t) + z(t)]_+ \quad (3)$$

where $f(t)$ is the firing rate in trial $t$ (we consider epochs of 1 second as “trials”), $z(t)$ is a diffusion variable, randomly and independently drawn from a normal distribution with mean 0 and variance $\sigma^2 = 0.01$ and $[x]_+$ denotes a linear-threshold function, $[x]_+ = x$ if $x \geq 0$ and 0 otherwise.
These random-walk neurons are clearly not action-value neurons. Nevertheless, we tested them using the analyses depicted in Fig. 1. To that goal, we randomly matched the trials in the simulation of the random-walk neurons (completely unrelated to the task) to the trials in the simulation depicted in Fig. 1A. Then, we considered the spike counts of the random-walk neurons in the last 20 trials of each of the four blocks in Fig. 1A (block being defined by the simulation of learning and is unrelated to the neural activity of the random-walk neurons). Surprisingly, when considering the top neuron in Fig. 2A and utilizing the same analysis as in Fig. 1B, we found that its spike count differed significantly between the first two blocks (p < 0.01, rank sum test) but not between the last two blocks (p = 0.28, rank sum test), similarly to the simulated action 1-value neuron of Fig. 1B (red). Similarly, the spike count of the bottom random-walk neuron matched that of a simulated action 2-value neuron (compare with the blue-labeled neuron in Fig. 1B; Fig. 2A).

Moreover, we regressed each vector of spike counts for 20,000 random-walk neurons on randomly matched estimated action-values from Fig. 1E and computed the t-values (Fig. 2B). This analysis erroneously classified 42% of these random-walk neurons as action-value neurons (see Fig. 2C). In particular, the top and bottom random-walk neurons of Fig. 2A were identified as action-value neurons for actions 1 and 2, respectively (squares in Fig. 2B).

To further quantify this result, we computed the fraction of random-walk neurons erroneously classified as action-value neurons as a function of the diffusion parameter $\sigma$ (Fig. 2D). When $\sigma=0$, the spike counts of the neurons in the different trials are independent and the number of random-walk neurons classified as action-value neurons is slightly less than 10%, the fraction expected by chance from a significance criterion of 5% and two statistical tests, corresponding to the two action-values. The larger the value of $\sigma$, the higher the probability that a random-walk
neuron will pass the selection criterion for at least one action-value and thus be erroneously identified as an action-value, state or policy neuron.

**Figure 2** Erroneous detection of action-value representation in random-walk neurons (A) Two example random-walk neurons that appear as if they represent action-values. The red (top) and blue (bottom) lines denote the estimated action-values 1 and 2, respectively that were depicted in Fig. 1B,C. Gray lines and gray dots denote the firing rates and the spike counts of two example random-walk neurons that were randomly assigned to this simulated session. Black horizontal lines denote the mean spike count in the last 20 trials of the block. Error bars denote the standard error of the mean. The two asterisks denote p<0.01 (rank sum test). (B) and (C) Population analysis. Each random-walk neuron was regressed on the two estimated action-values, as in Figs. 1D and 1E. Numbers and legend are the same as in Fig. 1D and 1E. The two random-walk neurons in (A) are denoted by squares in (B). Dashed lines in (B) at t=2 denote the significance boundaries. Dashed lines in (C) denote the naïve expected false positive rate from the significance threshold (see Materials and Methods). (D) Fraction of random-walk neurons classified as action-value neurons (red), and classified as state neurons (ΣQ) or policy neurons (ΔQ) (green) as a function of the magnitude of the diffusion parameter of random-walk (σ). Light red and light green are standard error of the mean. Dashed lines mark the results for σ=0.1, which is the value of the diffusion parameter used in (A)-(C). Initial firing rate for all neurons in the simulations is f(1) = 2.5Hz.
The excess action-value neurons in Fig. 2 emerged because the significance boundary in the statistical analysis was based on the assumption that the different trials are independent from each other. In the case of a regression of a random-walk process on an action-value related variable, this assumption is violated. The reason is that in this case, both predictor (action-value) and the dependent variable (spike count) slowly change over trials, the former because of the learning and the latter because of the random drift. As a result, the statistic, which relates these two signals, is correlated between trials, violating the independence-of-trials assumption of the test. Because of these dependencies, the expected variance of the statistic (be it average spike count in 20 trials or the regression coefficient), which is calculated under the independence-of-trials assumption, is an underestimate of the actual variance. Therefore, the fraction of random-walk neurons classified as action-value neurons increases with the magnitude of the diffusion, which is directly related to the magnitude of correlations between spike counts in proximate trials (Fig. 2D). The phenomenon of spurious significant correlations in time-series with temporal correlations has been described previously in the field of econometrics and a formal discussion of this issue can be found in (Granger & Newbold, 1974; Phillips, 1986).

**Is this confound relevant to the question of action-value representation in the striatum?**

**Is a random-walk process a good description of striatal neurons’ activity?**

The Gaussian random-walk process is just an example of a temporally correlated firing rate and we do not argue that the firing rates of striatal neurons follow a random-walk process. However, any other type of temporal correlations, e.g., oscillations or trends, will violate the independence-of-trials assumption, and may lead to the erroneous classification of neurons as representing action-values. Such temporal correlations can also emerge from stochastic learning. For example, in Fig. 2 – Figure supplement 1 we consider a model of operant leaning that is based on
covariance based synaptic plasticity (Loewenstein, 2008, 2010; Loewenstein & Seung, 2006; Neiman & Loewenstein, 2013) and competition (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). Because such plasticity results in slow changes in the firing rates of the neurons, applying the analysis of Fig. 1E to our simulations results in the erroneous classification of 43% of the simulated neurons as representing action-values. This is despite the fact that action-values are not computed as part of this learning, neither explicitly or implicitly.

Are temporal correlations in neural recordings sufficiently strong to affect the analysis?

To test the relevance of this confound to experimentally-recorded neural activity, we repeated the analysis of Fig. 2B,C on neurons recorded in two unrelated experiments: 89 neurons from extracellular recordings in the motor cortex of an awake monkey (Fig. 2 – Figure supplement 2A-B) and 39 auditory cortex neurons recorded intracellularly in anaesthetized rats (Fig. 2 – Figure supplement 2C-D; (Hershenhoren, Taaseh, Antunes, & Nelken, 2014)). We regressed the spike counts on randomly matched estimated action-values from Fig. 1E. In both cases we erroneously identified action-value representations (36% and 23%, respectively) in a fraction comparable to that reported in the striatum.

Strong temporal correlations in the striatum

To test the relevance of this confound to striatal neurons, we considered previous recordings from neurons in the nucleus accumbens (NAc) and ventral pallidum (VP) of rats in an operant learning experiment (Ito & Doya, 2009) and regressed their spike counts on simulated, unrelated action-values (using more blocks than in Fig. 1E, see Figure legend). Note that although the recordings were obtained during an operant learning task, the action-values that we used in the regression were obtained from simulated experiments and were completely unrelated to the true
experimental settings. Again, we erroneously identified a substantial fraction of neurons (43%) as representing action-values, a fraction comparable to that reported in the striatum (Fig. 2 – Figure supplement 3).

Haven't previous publications acknowledged this confound and successfully addressed it?

We conducted an extensive literature search to see whether previous studies have identified this confound and addressed it (see Materials and Methods). Two studies noted that processes such as slow drift in firing rate may violate the independence-of-trials assumption of the statistical tests and suggested unique methods to address this problem (H. Kim et al., 2013, 2009): one method (H. Kim et al., 2009) relied on permutation of the spike counts within a block (Fig. 2 – Figure supplement 4, see Materials and Methods) and another (H. Kim et al., 2013), relied on using spikes in previous trials as predictors (Fig. 2 – Figure supplement 5). However, both approaches fail to account for all temporal correlations and as a result, they still erroneously identify unrelated recorded and random-walk neurons as action-value neurons (Fig. 2 – Figure supplements 4,5). The failure of both these approaches stems from the fact that a complete model of the learning-independent temporal correlations is lacking. As a result, these methods are unable to remove all the temporal correlations from the vector of spike-counts.

Our literature search yielded four additional methods that have been used to identify action-value neurons. However, as depicted in Fig. 2 – Figure supplement 6 (corresponding to the analyses in (Ito & Doya, 2009; Samejima et al., 2005)), Fig. 2 – Figure supplement 7 (corresponding to the analysis in (Ito & Doya, 2015a)), Fig. 2 – Figure supplement 8 (corresponding to the analysis in (Wang et al., 2013)) and Fig. 2 – Figure supplement 9 (corresponding to a trial design experiment in (Fitzgerald et al., 2012)), all these additional methods erroneously identify neurons from unrelated recordings and random-walk neurons as action-value neurons in numbers.
comparable to those reported in the striatum (Fig. 2 – Figure supplements 6-9). The fMRI analysis in (Fitzgerald et al., 2012) focused on the difference between action-values rather than on the action-values themselves (see confound 2), and therefore we did not attempt to replicate it (and cannot attest to whether it is subject to the temporal correlations confound). We did, however, conduct the standard analysis on their unique experimental design - a trial-design experiment in which trials with different reward probabilities are randomly intermingled. Surprisingly, we erroneously detect action-value representation even when using this trial design (Fig. 2 – Figure supplement 9). This erroneous detection occurs because in this analysis the regression’s predictors are estimated action-values, which are temporally correlated. From this example it follows that even trial-design experiments may still be subject to the temporal correlations confound.

Some previous publications used more blocks. Shouldn’t adding blocks solve the problem?

In Figs. 1 and 2 we considered a learning task composed of four blocks with a mean length of 174 trials (standard deviation 43 trials). It is tempting to believe that experiments with more blocks and trials (e.g., (Ito & Doya, 2009; Wang et al., 2013)) will be immune to this confound. The intuition is that the larger the number of trials, the less likely it is that a neuron that is not modulated by action-value (e.g., a random-walk neuron) will have a large regression coefficient on one of the action-values. However, surprisingly, this intuition is wrong. In Fig. 2 – Figure supplement 10 we show that doubling the number of blocks, so that the original blocks are repeated twice, each time in a random order, does not decrease the fraction of neurons erroneously classified as representing action-values. For the case of random-walk neurons, it can be shown that, contrary to this intuition, the fraction of erroneously identified action-value neurons is expected to increase with the number of trials (Phillips, 1986). This is because the
expected variance of the regression coefficients under the null hypothesis is proportional to the inverse of the number of degrees of freedom, which increases with the number of trials. As a result, the threshold for classifying a regression coefficient as significant decreases with the number of trials.

**Possible solutions to the temporal correlations confound**

The temporal correlations confound has been acknowledged in the fMRI literature, and several methods have been suggested to address it, such as ‘prewhitening’, e.g., (Woolrich, Ripley, Brady, & Smith, 2001). However, these methods require prior knowledge, or an estimate of the predictor-independent temporal correlations. Both are impractical for the slow time-scale of learning and therefore are not applicable in the experiments we discussed.

Another suggestion is to assess the level of autocorrelations between trials in the data and to use it to predict the fraction of erroneous classification of action-value neurons. However, using such a measure is problematic in the context of action-value representation because the autocorrelations relevant for the temporal correlations confound are those associated with the time-scale relevant for learning, tens of trials. Computing such autocorrelations in experiments of a few hundreds of trials introduces substantial biases (Kohn, 2006; Newbold & Agiakloglou, 1993). Moreover, even when these autocorrelations are computed, it is not clear how they can be used to estimate the expected fraction of erroneously-classified action-value neurons.

Finally, it has been suggested that the temporal correlation confound can be addressed by using repeating blocks and removing neurons whose activity is significantly different in identical blocks (Asaad, Rainer, & Miller, 2000; Mansouri, Matsumoto, & Tanaka, 2006). We applied this method by applying a design in which the four blocks of Fig. 1 are repeated twice. However,
even when this method was applied, a significant number of neurons were erroneously classified as representing action-values (Materials and Methods).

Therefore, we propose two alternative approaches:

**Permutation analysis**

Trivially, an action-value neuron (or any task-related neuron) should be more strongly correlated with the action-values of the experimental session, in which the neuron was recorded than with action-values of other sessions (recorded in different days). We propose to use this requirement in a permutation test, as depicted in Fig. 3. We first consider the two simulated action-value neurons of Fig. 1B. For each of the two neurons, we computed the t-values of the regression coefficients of the spike counts on each of the estimated action-values in all possible sessions (see Materials and Methods). Fig. 3A depicts the two resulting distributions of t-values. As a result of the temporal correlations, the 5% significance boundaries (vertical dashed lines), which are defined to be exceeded by exactly 5% of t-values in each distribution, are substantially larger (in absolute value) than 2, the standard significance boundaries. We posit that the neuron is significantly correlated with an action-value if the t-value of the regression on the action-value (from the corresponding session) exceeds the significance boundaries derived from the permutation test.

Indeed, when considering the Top (red) simulated action 1-value neuron, we find that its spike count has a significant regression coefficient on the estimated $Q_1$ from its session (red arrow) but not on the estimated $Q_2$ (blue arrow). Importantly, because the significance boundary exceeds 2, this approach is less sensitive than the original one (Fig. 1) and indeed, the regression coefficients of the Bottom simulated neuron (blue) do not exceed the significance level (red and
blue arrows) and thus this analysis fails to identify it as an action-value neuron. Considering the population of simulated action-value neurons of Fig. 1, this analysis identified 29% of the action-value neurons of Fig. 1 as such (Fig. 3B, black), demonstrating that this analysis can identify action-value neurons. When considering the random-walk neurons (Fig. 2), this method classifies only approximately 10% of the random-walk neurons as action-value neurons, as predicted by chance. Similar results were obtained for the motor cortex and auditory cortex neurons (not shown).

Permutation analysis of basal ganglia neurons
Importantly, this permutation method can also be used to reanalyze the activity of previously-recorded neurons. To that goal, we considered the recordings reported in (Ito & Doya, 2009). The results of their model-free method (Fig. 2 – Figure supplement 6) imply that approximately 23% of the striatal neurons represent action-values at different phases of the experiment. As a first step, we estimated the action-values and regressed the spike counts in the different phases of the experiment on the estimated action-values, as in Fig. 1 (activity in each epoch is analyzed as if it is a different neuron; see Materials and Methods). The results of this analysis implied that 32% of the neurons represent action values (p<0.01) (Figure 3 – figure supplement 1). Next, we applied the permutation analysis. Remarkably, this analysis yielded that only 3.6% of the neurons have a significantly higher regression coefficient on an action-value from their session.
than on other action-values (Fig. 3C). Similar results were obtained when performing a similar model-free permutation analysis (regression of spike counts in the last 20 trials of the block on reward probabilities, not shown). These results raise the possibility that all or much of the apparent action-value representation in (Ito & Doya, 2009) is the result of the temporal correlations confound.

Figure 3 Permutation analysis (A) Red and blue correspond to red and blue - labeled neurons in Fig. 1B, respectively. Arrowheads denote the t-values from regressions on the estimated action values from the session in which the neuron was simulated (depicted in Fig. 1A). The red and blue histograms denote the t-values of the regressions of the spike-count on estimated action-values from different sessions in Fig. 1E (Materials and Methods). Dashed black lines denote the 5% significance boundary. We say that the regression coefficient of neural activity on an action-value is significant in this analysis if it exceeds these significance boundaries. Note that because of the temporal correlations, these boundaries are larger than ±2 (the significance boundaries in Figs. 1,2). According to this permutation test the red-labeled but not the blue-labeled neuron is classified as an action-value neuron (B) Fraction of neurons classified in each category using the permutation analysis for the action-value neurons (green, Fig. 1), random-walk neurons (yellow, Fig. 2). Dashed line denotes chance level for action-value 1 or 2 classification. Error bars denote the standard error of the mean. The test correctly
identifies 29% of actual action-value neurons as such, while classification of random walk neurons was at chance level. Analysis was done on 10,080 action-value neurons and 10,077 random-walk neurons from 504 simulated sessions (C) Light orange, fraction of basal ganglia neurons from (Ito & Doya, 2009) classified in each category when regressing the spike count of 214 basal ganglia neurons in three different experimental epochs on the estimated action-values associated with their experimental session. This analysis classified 32% of neurons as representing action-values. Dark orange, fraction of basal ganglia neurons classified in each category when applying the permutation analysis. This test classified 3.6% of neurons as representing action-value. Dashed line denotes significance level of p<0.01.

Trial-design experiments

Another way of overcoming the temporal correlations confound is to use a trial design in the experiment. The idea is to randomly mix the reward probabilities, rather than use blocks as in Fig. 1. For example, we propose the experimental design depicted in Fig. 4A. Each trial is presented in one of four clearly-marked contexts (color coded). The reward probabilities associated with the two actions are fixed within a context but differ between the contexts. Within each context the participant learns to prefer the action associated with a higher probability of reward. Naively, we can regress the spike counts on the action-values estimated from behavior, as in Fig. 1. However, because the estimated action-values are temporally correlated, this regression is still subject to the temporal correlations confound (Fig. 2 – Figure Supplement 9). Alternatively, we can regress the spike counts on the reward probabilities. If the contexts are randomly mixed, then by construction, the reward probabilities are temporally independent. These reward probabilities are the objective action-values. After learning, the subjective action-values are expected to converge to these reward probabilities. Therefore, the reward probabilities can be used as proxies for the subjective action-values after a sufficiently large number of trials and we can regress the spike counts on these reward probabilities. It is thus possible to conduct a
regression analysis on the spike counts at the end of the experiment, with reward probabilities as
predictors that do not violate the independence assumption.

To demonstrate this method, we simulated learning in a session composed of 400 trials, randomly divided into 4 different contexts (Fig. 4A). Learning followed the Q-learning equations (Eqs. 1 and 2), independently for each context. Next, we simulated action-value neurons, as in Fig. 1A, whose firing rate is a linear function of the action-value in the relevant context (dots in Fig. 4A). We regressed the spike counts of the neurons in the last 200 trials (approximately 50 trials in each context) on the corresponding reward probabilities (Fig. 4B). Indeed, 59% of the neurons were classified this way as action-value neurons (Fig. 4C 10% is chance level). By contrast, considering random-walk neurons, only 8.5% were erroneously classified as action-value neurons, a fraction expected by chance.

Figure 4 A possible solution for the temporal correlations confound that is based on trial design. (A) A Q-learning model was simulated in 1,000 sessions of 400 trials, where the original reward probabilities (same as in Fig. 1A) were associated with...
different cues and appeared randomly. Learning was done separately for each cue. Top panel: The first 20 trials in an example session. Background colors denote the reward probabilities in each trial. Black circles denote the learned value of action-value 1 in each trial. Top and bottom black lines denote choices of action 1 and 2, respectively. Long and short lines denote rewarded and unrewarded trials, respectively. Bottom panels: Two examples of the grouping of trials with the same reward probabilities to show the continuity in learning. Note that the action-value changes only when action 1 is chosen because it is the action-value associated with action 1. (B) and (C) population analysis for action-value neurons. 20,000 action-value neurons were simulated from the model in (A), similarly to the action-value neurons in Fig. 1. For each neuron, the spike-counts in the last 200 trials of the session were regressed on the reward probabilities (see Materials and Methods). Legend is the same in Figs. 1D-E. Dashed lines in (B) at t=2 denote the significance boundaries. Dashed lines in (C) denote the naïve expected false positive rate from the significance threshold (see Materials and Methods). This analysis correctly identifies 59% of action-value neurons as such. (D) and (E) population analysis for random-walk neurons. 20,000 Random-walk neurons were simulated as in Fig. 2. Same regression analysis as in (B) and (C). Only 8.5% of the neurons were erroneously classified as representing action-values (9.5% chance level).

Three previous studies used related trial-designs to search for action-value representation in the striatum (Cai et al., 2011; Fitzgerald et al., 2012; S. Kim et al., 2012). However, in two of them (Cai et al., 2011; S. Kim et al., 2012) the reward probabilities were explicitly cued and therefore their results can be interpreted in the framework of cue-values and not action-values (Padoa-Schioppa, 2011). Moreover, all these studies focused on significant neural modulation by both action-values or by their difference, analyses that support state or policy representations (Ito & Doya, 2015a). As discussed in details in the next section, policy representation can emerge without action-value representation (Darshan, Leblois, & Hansel, 2014; Fiete, Fee, & Seung, 2007; Fremaux, Sprekeler, & Gerstner, 2010; Loewenstein, 2008, 2010; Loewenstein & Seung, 2006; Neiman & Loewenstein, 2013; Seung, 2003; Urbanczik & Senn, 2009). Therefore, the results reported in (Cai et al., 2011; Fitzgerald et al., 2012; S. Kim et al., 2012) cannot be taken as evidence for action-value representation in the striatum.

**Confound 2 – correlated decision variables**
In the previous sections we demonstrated that irrelevant temporal correlations may lead to the erroneous identification of neurons as representing action-values, even if their activity is task-independent. Here we address an unrelated confound. We show that neurons that encode different decision variables, in particular policy, may be erroneously identified as representing action-values. For clarity, we will commence by discussing this caveat independently of the temporal correlations confound. Specifically, we show that neurons whose firing rate encodes the policy (probability of choice) may be erroneously identified as representing action-values, even when this policy emerged in the absence of any implicit or explicit action-value representation. We will conclude by discussing a possible solution that addresses this and the temporal correlations confounds.

**Policy without action-value representation**

It is well-known that operant learning can occur in the absence of any value computation, e.g., as a result of direct-policy learning (Mongillo, Shteingart, & Loewenstein, 2014). Several studies have shown that reward-modulated synaptic plasticity can implement direct policy reinforcement learning (Darshan et al., 2014; Fiete et al., 2007; Fremaux et al., 2010; Loewenstein, 2008, 2010; Loewenstein & Seung, 2006; Neiman & Loewenstein, 2013; Seung, 2003; Urbanczik & Senn, 2009).

For concreteness, we consider a particular reinforcement learning algorithm, in which the probability of choice $\Pr(a(t) = 1)$ is determined by a single variable $W$ that is learned in accordance to REINFORCE learning algorithm (Williams, 1992): $\Pr(a(t) = 1) = \frac{1}{1 + e^{-W(t)}}$

where $\Delta W(t) = \alpha \cdot (2 \cdot R(t) - 1) \cdot (a(t) - \Pr(a(t) = 1))$, where $\alpha$ is the learning rate, $R(t)$ is the binary reward in trial $t$ and $a(t)$ is a binary variable indicating whether action 1 was chosen
in trial $t$; in our simulations $W(t = 1) = 0$, $\alpha = 0.17$. For biological implementation of this algorithm see (Loewenstein, 2010; Seung, 2003).

We tested this model in the schedule of Fig. 1. (Fig. 5A). As expected, the model learned to prefer the action associated with a higher probability of reward, completing the four blocks within 228 trials on average (standard deviation 62 trials).

**Spike count of neurons representing policy are correlated with estimated $\Delta Q$**

Despite the fact that the learning was value-independent, we can still fit a Q-learning model to the behavior, extract best-fit model parameters and compute action-values (see also Fig. 2 – Figure supplement 1). The computed action-values are presented in Fig. 5B. Note that according to Eq. (2), the probability of choice is a monotonic function of the difference between $Q_1$ and $Q_2$. Therefore, we expected that the probability of choice will be correlated with the computed $Q_1$ and $Q_2$, with opposite signs (Fig. 5C).

We simulated policy neurons as Poisson neurons whose firing rate is a linear function of the policy $Pr(a)$ (Materials and Methods). Next, we regressed the spike counts of these neurons on the two action-values that were computed from behavior (same as in Figs. 1D,E, 2B,C, Fig. 2 – Figure supplement 1C,D, – Figure supplement 2B,D, – Figure supplement 3). Indeed, as expected, 14% of the neurons were significantly correlated with both action values with opposite signs (chance level for each action value is 5%, naïve chance level for both with opposite signs is 0.125%, see Materials and Methods), as depicted in Fig. 5D,E. These results demonstrate that neurons representing value-independent policy can be classified as representing $\Delta Q$. 
Figure 5 Erroneous detection of action-value representation in policy neurons. (A) Behavior of model in example session, same as in Fig. 1A for the direct policy model. (B) Red and blue lines denote "action-values" 1 and 2 respectively, calculated from the choices and rewards in (A). Note that the model learned without any explicit or implicit calculation of action-values. The extraction of action-values in (B) is based on the fitting of Eq. (1) to the learning behavior. (C) Strong correlation between policy from the direct policy algorithm depicted in (A) and action-values extracted by fitting Eq. (1) to behavior. The three panels depict probability of choice as a function of the difference between the calculated action-values (left), "$Q_1$" (center) and "$Q_2$" (right). This correlation can cause policy neurons to be erroneously classified as representing action-values (D) and (E) Population analysis, same as in Figs. 1D and 1E for the policy neurons. Legend and number of neurons are also as in Figs. 1D and 1E. Dashed lines in (B) at t=2 denote the significance boundaries. Dashed lines in (E) denote the naïve expected false positive rate from the significance threshold (see Materials and Methods).
Neurons representing policy may be erroneously classified as action-value neurons

Surprisingly however, 38% of policy neurons were significantly correlated with exactly one estimated action-value, and therefore would have been classified as action-value neurons in the standard method of analysis (10% chance level).

To understand why this erroneous classification emerged, we note that a neuron is classified as representing an action-value if its spike count is significantly correlated with one of the action values, but not with the other. The confound that led to the classification of policy neurons as representing action-values is that a lack of statistically significant correlation is erroneously taken to imply lack of correlation. All policy neurons are modulated by the probability of choice, a variable that is correlated with the difference in the two action-values. Therefore, this probability of choice is expected to be correlated with both action-values, with opposite signs.

However, because the neurons are Poisson, the spike count of the neurons is a noisy estimate of the probability of choice. As a result, in most cases (86%), the regression coefficients do not cross the significance threshold for both action-values. More often (38%), only one of them crosses the significance threshold, resulting in an erroneous identification of the neurons as representing action values.

Is this confound relevant to the question of action-value representation in the striatum?

If choice is included as a predictor, is policy representation still a relevant confound?

It is common, (although not ubiquitous) to attempt to differentiate action-value representation from choice representation by including choice as another regressor in the regression model (Cai et al., 2011; Fitzgerald et al., 2012; Funamizu et al., 2015; Her et al., 2016; Ito & Doya, 2015a, 2015b, H. Kim et al., 2013, 2009; S. Kim et al., 2012; Lau & Glimcher, 2008). Such analyses
may be expected to exclude policy neurons, whose firing rate is highly correlated with choice, from being classified as action-value neurons. However, repeating this analysis for the policy neurons of Fig. 5, we still classify 36% of policy neurons as action-value neurons (Fig. 5 – Figure supplement 1A).

An alternative approach has been to consider only those neurons whose spike count is not significantly correlated with choice (Stalnaker et al., 2010; Wunderlich et al., 2009). Repeating this analysis for Fig. 5 policy neurons, we still find that 24% of the neurons are erroneously classified as action-value neurons (8% are classified as policy neurons).

Is this confound the result of an analysis that is biased against policy representation?

The analysis depicted in Figs. 1D,E, 2B,C, 4B-E, and 5D,E is biased towards classifying neurons as action-value neurons, at the expense of state or policy neurons, as noted by (Wang et al., 2013). This is because action-value classification is based on a single significant regression coefficient whereas policy or state classification requires two significant regression coefficients. Therefore, (Wang et al., 2013) have proposed an alternative approach. First, compute the statistical significance of the whole regression model for each neuron (using f-value). Then, classify those significant neurons according to the t-value of regression coefficients on the two action-values (Fig. 5 – Figure supplement 1B). Applying this analysis to the policy neurons of Fig. 5 with a detection threshold of 5% we find that indeed, this method is useful in detecting which decision variables are more frequently represented (its major use in (Wang et al., 2013)): 25% of the neurons are indeed classified as representing policy (1.25% expected by chance). Nevertheless, 12% of the neurons are still classified as action-value neurons (2.5% expected by chance; Fig. 5 – Figure supplement 1B).
In many cases, the term action-value was used, while the reported results were equally consistent with other decision variables. In some cases, significant correlation with both action-values (with opposite signs) or significant correlation with the difference between the action-values was used as evidence for ‘action-value representations’ (Fitzgerald et al., 2012; Guitart-Masip et al., 2012; S. Kim et al., 2012; Y. B. Kim et al., 2007; Stalnaker et al., 2010). Similarly, other papers did not distinguish between neurons whose activity is significantly correlated with one action-value and those whose activity is correlated with both action-values (Funamizu et al., 2015; Her et al., 2016; H. Kim et al., 2013, 2009). Finally, one study used a concurrent variable-interval schedule, in which the magnitudes of rewards associated with each action were fully correlated (Lau & Glimcher, 2008). In such a design, the two probabilities of reward depend on past choices and therefore, the values associated with the actions change trial-by-trial and are, in general, correlated.

A possible solution to the policy confound

The policy confound emerged because policy and action-values are correlated. To distinguish between the two possible representations, we should seek a variable that is correlated with the action-value but uncorrelated with the policy. Consider the sum of the two action-values. It is easy to see that \( \text{Corr}(Q_1 + Q_2, Q_1 - Q_2) \propto \text{Var}(Q_1) - \text{Var}(Q_2) \). Therefore, if the variances of the two action-values are equal, their sum is uncorrelated with their difference. An action-value neuron is expected to be correlated with the sum of action-values. By contrast, a policy neuron, modulated by the difference in action-values is expected to be uncorrelated with this sum.
We repeated the simulations of Fig. 4 (which addresses the temporal correlations confound), considering three types of neurons: action-value neurons (of Fig. 1), random-walk neurons (of Fig. 2), and policy neurons (of Fig. 5). As in Fig. 4, we considered the spike counts of the three types of neurons in the last 200 trials of the session, but now we regressed them on the sum of reward probabilities (state; in this experimental design the reward probabilities are also the objective action-values, which the subject learns). We found that only 4.5% and 6% of the random-walk and policy neurons, respectively, were significantly correlated with the sum of reward probabilities (5% chance). By contrast, 47% of the action-value neurons were significantly correlated with this sum.

This method is able to distinguish between policy and action-value representations. However, it will fail in the case of state representation because both state and action-values are correlated with the sum of probabilities of reward. To dissociate between state and action-value representations, we can consider the difference in reward probabilities as this difference is correlated with the action-value but is uncorrelated with the state. Regressing the spike count on both the sum and difference in the probabilities of reward, a random-walk neuron is expected to be correlated with none, a policy neuron is expected to be correlated only with the difference, whereas an action-value neuron is expected to be correlated with both (this analysis is inspired by Fig. S8b in (Wang et al., 2013) in which the predictors in the regression model were policy and state). This is depicted in Fig. 6A,C,E, where we report the t-value of these correlations. We now classify a neuron that passes both significance tests as an action-value neuron. Indeed, for a significance threshold of $p<0.05$ (for each test), Only 0.2% of the random-walk neurons and 5% of the policy neurons were classified as action-value neurons. By contrast, 32% of the action-value neurons were classified as such. Note that in this analysis only when more than 5% of the
neurons are classified as action-value neurons we have support for the hypothesis that there is action-value rather than policy or state representation.

A word of caution is that the analysis should be performed only after the learning converges. This is because stochastic fluctuations in the learning process may be reflected in the activities of neurons representing decision-related variables. As a result, policy or state-representing neurons may appear correlated with the orthogonal variables. For the same reason, any block-related heterogeneity in neural activity could result in this confound (O’Doherty, 2014).

**Figure 6** Possible solution for the policy and state confounds. (A) The Q-learning behavioral model (Eqs. 1 and 2) was simulated in 1,000 sessions of 400 trials each, where the reward probabilities were associated with different cues and appeared randomly, as in Fig. 4. Learning occurred separately for each cue. In each session 20 action-value neurons, whose firing rate is proportional to the action-values (as in Fig. 1) were simulated. For each neuron, the spike-counts in the last 200 trials of each session were regressed on the sum of the reward probabilities ($\Sigma Q$; state) and the difference of the reward probabilities ($\Delta Q$; policy, see Materials and Methods). Each dot denotes the t-values of the two regression coefficients of each of 500 example neurons. Dashed lines at $t=2$ denote the 5% significance boundaries of the regression coefficients. Neurons that had significant regression coefficients with both policy and state were identified as action-value neurons. Colors as in Fig. 1D. (B) Population analysis revealed that 32% of the action-value
neurons were identified as such. Error bars are the standard error of the mean. Dashed black line denotes the expected false positive rate from randomly modulated neurons. Dashed gray line denotes the expected false positive rate from policy or state neurons (see Materials and Methods) (C) Same as in (A) with random-walk neurons, numbers are as in Fig. 2. (D) Population analysis revealed that less than 1% of the random-walk neurons were erroneously classified as representing action-values. (E-F) To test the policy neurons, we simulated a direct-policy learning algorithm (as in Fig. 5) in the same sessions as in (A-D). Learning occurred separately for each cue. In each session 20 policy neurons, whose firing rate is proportional to the probability of choice (as in Fig. 5) were simulated. As in (A-D), the spike-counts in the last 200 trials of each session were regressed on the sum and difference of the reward probabilities. (E) each dot denotes the t-values of the two regression coefficients of a single neuron out of 500 example neurons. (F) Population analysis. As expected, only 5% of the policy neurons were erroneously classified as representing action-values.

To conclude, it is worthwhile repeating the key features of the analysis method proposed in this section:

1) Trial design is necessary because otherwise temporal correlations in spike count may inflate the fraction of neurons that pass the significance tests.

2) Regression should be performed on reward probabilities (i.e., the objective action-values) and not on estimated action-values. The reason is that because the estimated action-values evolve over time, this trial design does not eliminate all temporal correlations between them (Fig. 2 – Figure supplement 9).

3) Reward probabilities associated with the two actions should be chosen such that their variances should be equal. Otherwise policy or state neurons may be erroneously classified as action-value neurons.

Discussion

In this paper, we performed a systematic literature search to discern the methods that have been previously used to infer the representation of action-values in the striatum. We showed that none of these methods overcome two critical confounds: (1) neurons with temporal correlations in
their firing rates may be erroneously classified as representing action-values and (2) neurons whose activity co-varies with other decision variables, such as policy, may also be erroneously classified as representing action-values. Finally, we discuss possible experiments and analyses that can address the question of whether neurons encode action-values.

Temporal correlations and action-value representations

It is well known in statistics that the regression coefficient between two independent slowly-changing variables is on average larger (in absolute value) than this coefficient when the series are devoid of a temporal structure. If these temporal correlations are overlooked, the probability of a false-positive is underestimated (Granger & Newbold, 1974). When searching for action-value representation in a block design, then by construction, there are positive correlations in the predictor (action-values). Positive temporal correlations in the dependent variable (neural activity) will result in an inflation of the false-positive observations, compared with the naïve expectation.

This confound occurs only when there are temporal correlations in both the predictor and the dependent variable. In a trial design, in which the predictor is chosen independently in each trial and thus has no temporal structure, we do not expect this confound. However, when studying incremental learning, it is difficult to randomize the predictor in each trial, making the task of identifying neural correlates of learning, and specifically action-values, challenging. With respect to the dependent variable (neural activity), temporal correlations in BOLD signal and their consequences have been discussed (Arbabshirani et al., 2014; Woolrich et al., 2001). Considering electrophysiological recordings, there have been attempts to remove these correlations, e.g., using previous spike counts as predictors (H. Kim et al., 2013). However, these are not sufficient because they are unable to remove all task-independent temporal correlations.
(see also Figs. 2 – Figure supplements 4-10). When repeating these analyses, we erroneously classified a fraction of action-value neurons that is comparable to that reported in the striatum. However, the probability of a false-positive identification of a neuron as representing action-value depends on the magnitude and type of temporal correlations in the neural activity. Therefore, we cannot predict the fraction of erroneously classified neurons expected in various experimental settings and brain areas and this fraction may be even larger than the one reported here.

One may argue that the fact that action-value representations are reported mostly in a specific brain area, namely the striatum, is an indication that their identification there is not a result of the temporal correlations confound. However, because different brain regions are characterized by different spiking statistics, we expect different levels of erroneous identification of action-value neurons in different parts of the brain and in different experimental settings. Indeed, the fraction of erroneously identified action-value neurons differed between the auditory and motor cortices (compare B and D within Fig. 2 – Figure supplement 2). Furthermore, many studies reported action-value representation outside of the striatum, in brain areas including the supplementary motor area and presupplementary eye fields (Wunderlich et al., 2009), the substantia nigra/ventral tegmental area (Guitart-Masip et al., 2012) and ventromedial prefrontal cortex, insula and thalamus (Fitzgerald et al., 2012).

Considering the ventral striatum, our analysis on recordings from (Ito & Doya, 2009) indicates that the identification of action-value representations there may have been erroneous, resulting from temporally correlated firing rates (Figs. 3 and Fig. 2 – Figure supplement 3). It should be noted that the fraction of action-value neurons reported in (Ito & Doya, 2009) is low relative to other publications, a difference that has been attributed to the location of the recording in the
striatum (ventral as opposed to dorsal). It would be interesting to apply this method to other
striatal recordings (Ito & Doya, 2015a; Samejima et al., 2005; Wang et al., 2013). We were
unable to directly analyze these recordings from the dorsal striatum because relevant raw data is
not publicly available. However, previous studies have reported that the firing rates of dorsal-
istrial neurons change slowly over time (Gouvea et al., 2015; Mello, Soares, & Paton, 2015). As
a result, identification of apparent action-value representation in dorsal-striatal neurons may also
be the result of this confound.

Temporal correlations naturally emerge in experiments composed of multiple trials. Participants
become satiated, bored, tired, etc., which may affect neuronal activity. In particular, learning in
operant tasks is associated, by construction, with variables that are temporally correlated. If
neural activity is correlated with performance (e.g., accumulated rewards in the last several
trials) then it is expected to have temporal correlations, which may lead to an erroneous
classification of the neurons as representing action-values.

**Temporal correlations – beyond action-value representation**

Action-values are not the only example of slowly-changing variables. Any variable associated
with incremental learning, motivation or satiation is expected to be temporally correlated. Even
'benign' behavioral variables, such as the location of the animal or the activation of different
muscles may change at relatively long time-scales. When recording neural activity related to
these variables, any temporal correlations in the neural recording, be it in fMRI,
electrophysiology or calcium imaging may result in an erroneous identification of correlates of
these behavioral variables because of the temporal correlation confound.
In general, the temporal correlation confound can be addressed by using the permutation analysis of Fig. 3, which can provide strong support to the claim that the activity of a particular neuron or voxel co-varies with the behavioral variable. Therefore, the permutation test is a general solution for scientists studying slow processes such as learning. More challenging, however, is precisely identifying what the activity of the neuron represents (for example an action-value or policy). There are no easy solutions to this problem and therefore caution should be applied when interpreting the data.

Correlated decision variables

Another difficulty in identifying action-value neurons is that they are correlated with other decision variables such as policy, state or chosen-value. Therefore, finding a neuron that is significantly correlated with an action-value could be the byproduct of its being modulated by other decision variables, in particular policy. The problem is exacerbated by the fact that standard analyses (e.g., Fig. 1D-E) are biased towards classifying neurons as representing action-values at the expense of policy or state. However, because of the correlation between policy and action-value, even unbiased analyses may erroneously identify a significant fraction of neurons as representing action-values (Fig. 5 – Figure supplement 1B).

Differentiating action-value from other decision variables

As shown in Fig. 6, policy representation can be ruled out by finding a representation that is orthogonal to policy, namely state representation. This solution leads us, however, to a serious conceptual issue. All analyses discussed so far are based on significance tests: we divide the space of hypothesis into the "scientific claim" (e.g., neurons represent action-values) and the null
hypothesis (e.g., neural activity is independent of the task). An observation that is not consistent
with the null hypothesis is taken to support the alternative hypothesis.

The problem we faced with correlated variables is that the null hypothesis and the "scientific
claim" were not complementary. A neuron that represents policy is expected to be inconsistent
with the null hypothesis that neural activity is independent of the task but it is not an action-value
neuron. The solution proposed was to devise a statistical test that seeks to identify a
representation that is correlated with action-value and is orthogonal to the policy hypothesis, in
order to also rule out a policy representation.

However, this does not rule out other decision-related representations. A "pure" action-value
neuron is modulated only by $Q_1$ or by $Q_2$. A "pure" policy neuron is modulated exactly by
$Q_1 - Q_2$. More generally, we may want to consider the hypotheses that the neuron is modulated
by a different combination of the action values, $a \cdot Q_1 + b \cdot Q_2$, where $a$ and $b$ are parameters.

For every such set of parameters $a$ and $b$ we can devise a statistical test to reject this hypothesis
by considering the direction that is orthogonal to the vector $(a, b)$. In principle, this procedure
should be repeated for every pair of parameters $a$ and $b$ that in not consistent with the action-
value hypothesis.

Put differently, in order to find neurons that represent action-values, we first need to define the
set of parameters $a$ and $b$ such that a neuron whose activity is modulated by $a \cdot Q_1 + b \cdot Q_2$ will
be considered as representing an action-value. Only after this (arbitrary) definition is given, can
we construct a set of statistical tests that will rule out the competing hypotheses, namely will rule
out all values of $a$ and $b$ that are not in this set. The analysis of Fig. 6 implicitly defined the set of
$a$ and $b$ such that $a \neq b$ and $a \neq -b$ as the set of parameters that defines action-value
representations. In practice, it is already very challenging to identify action-values using the procedure of Fig. 6 and going beyond it seems impractical. Therefore, studying the distribution of t-values across the population of neurons may be more useful when studying decision-variables representations than asking questions about the significance of individual neurons.

Importantly, the regression models described in this paper allow us to investigate only some types of representations, namely, linear combinations of the two action-values. However, value representations in learning models may fall outside of this regime. It has been suggested that in decision making, the ratio of action-values is calculated (Worthy, Maddox, & Markman, 2008), or that participants compute, for each action, the probability that it is associated with the highest value (Morris, Dezfooli, Griffiths, & Balleine, 2014). Our proposed solution cannot support or refute these alternative hypotheses. If these are taken as additional alternative hypotheses, a neuron should be classified as representing an action-value if its activity is also significantly modulated in the directions that are correlated with action-value and are orthogonal to these hypotheses. Clearly, it is never possible to construct an analysis that can rule out all possible alternatives.

We believe that the confounds that we described have been overlooked because the null hypothesis in the significance tests was not made explicit. As a result, the complementary hypothesis was not explicitly described and the conclusions drawn from rejecting the null hypothesis were too specific. That is, alternative plausible interpretations were ignored. It is important, therefore, to keep the alternative hypotheses explicit when analyzing the data, be it using significance tests or other methods, such as model comparison (Ito & Doya, 2015b).

Are action-value representations a necessary part of decision making?
One may argue that the question of whether neurons represent action-value, policy, state or some other correlated variable is not an interesting question. This is because all these correlated decision variables implicitly encode action-values. Even direct-policy models can be taken to implicitly encode action-values because policy is correlated with the difference between the action-values. However, we believe that the difference between action-value representation and representation of other variables is an important one, because it centers on the question of the computational model that underlies decision making in these tasks. Specifically, the implication of a finding that a population of neurons represent action-values is not that these neurons are involved somehow in decision making. Rather, we interpret this finding as supporting the hypothesis that action-values are explicitly computed in the brain, and that these action-values play a specific role in the decision making process. However, if the results are also consistent with various alternative computational models then this is not the case. Some consider action-value computation to be a necessary part of decision making. By contrast, however, we presented here two models of learning and decision making that do not entail this computation (Fig. 2 – Fig. supplement 1, Fig. 5). Other examples are discussed in (Mongillo et al., 2014; Shteingart & Loewenstein, 2014) and references therein.

Other indications for action-value representation

Several trial-design experiments have associated cues with upcoming rewards and reported representations of expected reward, the upcoming action, or the interaction of action and reward (H. C. Cromwell & Schultz, 2003; Howard C. Cromwell, Hassani, & Schultz, 2005; Hassani, Cromwell, & Schultz, 2001; Hori, Minamimoto, & Kimura, 2009; Kawagoe, Takikawa, & Hikosaka, 1998; Pasquereau et al., 2007). Another trial-design experiment reported representation of offer-value and chosen-value in the orbitofrontal cortex (Padoa-Schioppa &
Assad, 2006). While such studies do not provide direct evidence for action-value representation, they do provide evidence for representation of closely-related decision variables (but see O’Doherty, 2014).

The involvement the basal ganglia in general and the striatum in particular in operant learning, planning and decision-making is well documented (Ding & Gold, 2010; McDonald & White, 1993; O’Doherty et al., 2004; Palminteri et al., 2012; Schultz, 2015; Tai et al., 2012; Thorn, Atallah, Howe, & Graybiel, 2010; Yarom & Cohen, 2011). However, there are alternatives to the possibility that the firing rate of striatal neurons represents action-values. First, as discussed above, learning and decision making do not entail action-value representation. Second, it is possible that action-value is represented elsewhere in the brain. Finally, it is also possible that the striatum plays an essential role in learning, but that the representation of decision variables there is distributed and neural activity of single neurons could reflect a complex combination of value-related features, rather than "pure" decision variables. Such complex representations are typically found in artificial neural networks (Yamins & DiCarlo, 2016).

**Action-value representation in the striatum requires further evidence**

Considering the literature, both confounds have been partially acknowledged. Moreover, there have been some attempts to address them. However, as discussed above, even when these confounds were acknowledged and solutions were proposed, these solutions do not prevent the erroneous identification of action-value representation (see Fig. 2 – Figure supplements 4,5,10, Fig. 5 – Figure supplement 1). We therefore conclude that to the best of our knowledge, all studies that have claimed to provide direct evidence that neuronal activity in the striatum is specifically modulated by action-value were either susceptible to the temporal correlations confound (Funamizu et al., 2015; Ito & Doya, 2009, 2015a, 2015b, H. Kim et al., 2013, 2009;
Lau & Glimcher, 2008; Samejima et al., 2005; Wang et al., 2013), or reported results in a manner indistinguishable from policy (Cai et al., 2011; Fitzgerald et al., 2012; Funamizu et al., 2015; Guitart-Masip et al., 2012; Her et al., 2016; H. Kim et al., 2013, 2009; S. Kim et al., 2012; Y. B. Kim et al., 2007; Stalnaker et al., 2010; Wunderlich et al., 2009). Many other studies differentiated action-value and policy, but were subject to the second confound (Ito & Doya, 2009, 2015a, 2015b; Lau & Glimcher, 2008; Samejima et al., 2005). Furthermore, it should be noted that not all studies investigating the relation between striatal activity and action-value representation have reported positive results. Several studies have reported that striatal activity is more consistent with direct policy learning than with action-value learning (FitzGerald, Schwartenbeck, & Dolan, 2014; Li & Daw, 2011) and one noted that lesions to the dorsal striatum do not impair action-value learning (Vo, Rutledge, Chatterjee, & Kable, 2014).

Finally, we would like to emphasize that we do not claim that there is no representation of action-value in the striatum. Rather, our results show that special caution should be applied when relating neural activity to reinforcement-learning related variables. Therefore, the prevailing belief that neurons in the striatum represent action-values must await further tests that address the confounds discussed in this paper.

Materials and Methods

Literature search

In order to thoroughly examine the finding of action-value neurons in the striatum, we conducted a literature search to find all the different approaches used to identify action-value representation in the striatum and see whether they are subject to at least one of the two confounds we described here.
Key words “action-value” and “striatum” were searched for in Web-of-Knowledge, Pubmed and Google Scholar, returning 43, 21 and 980 results, respectively. In the first screening stage, we excluded all publications that did not report new experimental results (e.g., reviews and theoretical papers), focused on other brain regions, or did not address value-representation or learning. In the remaining publications, the abstract of the publication was read and the body of the article was searched for “action-value” and “striatum”. After this step, articles in which it was possible to find description of action-value representation in the striatum were read thoroughly. The search included PhD theses, but none were found to report new relevant data, not found in papers. We identified 22 papers that directly related neural activity in the striatum to action-values. These papers included reports of single-unit recordings, functional magnetic resonance imaging (fMRI) experiments and manipulation of striatal activity.

Of these, 2 papers have used the term action-value to refer to the value of the chosen action (also known as chosen-value) (Day, Jones, & Carelli, 2011; Seo, Lee, & Averbeck, 2012) and therefore we do not discuss them.

An additional study (Pasquereau et al., 2007) used the expected reward and the chosen action as predictors of the neuronal activity and found neurons that were modulated by the expected reward, the chosen action and their interaction. The authors did not claim that these neurons represent action-values, but it is possible that these neurons were modulated by the values of specific actions. However, the representation of the value of the action when the action is not chosen is a crucial part of action-value representation which differentiates it from the representation of expected reward, and the values of the actions when they were not chosen were not analyzed in this study. Therefore, the results of this study cannot be taken as an indication for action-value representation, rather than other decision variables.
A second group of 11 papers did not distinguish between action-value and policy representations (Cai et al., 2011; Funamizu et al., 2015; Her et al., 2016; H. Kim et al., 2013, 2009; Wunderlich et al., 2009), or reported policy representation (Fitzgerald et al., 2012; Guitart-Masip et al., 2012; S. Kim et al., 2012; Y. B. Kim et al., 2007; Stalnaker et al., 2010) in the striatum and therefore their findings do not necessarily imply action-value representation, rather than policy representation in the striatum (see confound 2).

In 2 additional papers, it was shown that the activation of striatal neurons changes animals’ behavior, and the results were interpreted in the action-value framework (Lee et al., 2015; Tai et al., 2012). However, a change in policy does not entail an action-value representation (see, for example, Figs. 5 and Fig. 2 – Figure supplement 1). Therefore, these papers were not taken as strong support to the striatal action-value representation hypothesis.

Finally, 6 papers correlated action-values, separately from other decision variables, with neuronal activity in the striatum (Ito & Doya, 2009, 2015a, 2015b; Lau & Glimcher, 2008; Samejima et al., 2005; Wang et al., 2013). All of them used electrophysiological recordings of single units in the striatum. From these papers, only one utilized an analysis which is not biased towards identifying action-value neurons at the expense of policy and state neurons (Wang et al., 2013). All papers used block-design experiments where action-values are temporally correlated.

Taken together, we concluded that previous reports on action-value representation in the striatum could reflect the representation of other decision variables or temporal correlations in the spike count that are not related to action-value learning.

The action-value neurons model (Fig. 1)
To model neurons whose firing rate is modulated by an action-value (Fig. 1), we considered neurons whose firing rate changes according to:

\[ f(t) = B + K \cdot r \cdot (Q_i(t) - 0.5) \]  (4)

Where \( f(t) \) is the firing rate in trial \( t \), \( B = 2.5\)Hz is the baseline firing rate, \( Q_i(t) \) is the action-value associated with one of the targets \( i \in \{1,2\} \), \( K = 2.35\)Hz is the maximal modulation and \( r \) denotes the neuron-specific level of modulation, drawn from a uniform distribution, \( r \sim U[-1,1] \). The spike count in a trial was drawn from a Poisson distribution, assuming a 1 sec-long trial.

*The policy neurons model* (Fig. 5)

To model neurons whose firing rate is modulated by a policy, we considered neurons whose firing rate changes according to:

\[ f(t) = B + K \cdot r \cdot (\Pr(a(t) = 1) - 0.5) \]  (5)

Where \( f(t) \) is the firing rate in trial \( t \), \( B = 2.5\)Hz is the baseline firing rate, \( \Pr(a(t) = 1) \) is the probability of choosing action 1 in trial \( t \) that changes in accordance with REINFORCE (Williams, 1992) (see also Fig. 5 and corresponding text). \( K = 3\)Hz is the maximal modulation and \( r \) denotes the neuron-specific level of modulation, drawn from a uniform distribution, \( r \sim U[-1,1] \). The spike count in a trial was drawn from a Poisson distribution, assuming a 1 sec-long trial.

*The covariance neurons model* (Fig. 2 – Figure supplement 1)

In the covariance based plasticity model the decision-making network is composed of two populations of Poisson neurons: each neuron is characterized by its firing rate and the spike count of a neuron in a trial (1 sec) is randomly drawn from a Poisson distribution. The chosen
action corresponds to the population that fires more spikes in a trial (Loewenstein, 2010; Loewenstein & Seung, 2006). At the end of the trial, the firing rate of each of the neurons (in the two population) is updated according to \( f(t + 1) = f(t) + \eta \cdot R(t) \cdot (s(t) - f(t)) \), where \( f(t) \) is the firing rate in trial \( t \), \( \eta = 0.07 \) is the learning rate, \( R(t) \) is the reward delivered in trial \( t \) \((R(t) \in \{0,1\} \) in our simulations\) and \( s(t) \) is the measured (realized) firing rate in that trial, that is the spike count in the trial. The initial firing rate of all simulated neurons is 2.5Hz. The network model was tested in the operant learning task of Fig. 1. A session was terminated (without further analysis) if the model was not able to choose the better option more than 14/20 times for at least 200 trials. This occurred on 20% of the sessions. We simulated two populations of 1,000 neurons in 500 successful sessions. Simulated neurons were excluded due to low spike rate if the mean spike count was lower than 1 for all blocks. This occurred to 0.03% of the neurons. Note that because on average, the empirical firing rate is equal to the true firing rate, \( f(t) = \langle s(t) \rangle \), changes in the firing rate are driven, on average, by the covariance of reward and the empirical firing rate: \( \langle \Delta f(t) \rangle \equiv \langle f(t + 1) - f(t) \rangle = \eta \cdot \text{cov}(R(t), s(t)) \) (Loewenstein & Seung, 2006). The estimated Q-values in Fig. 2 – Figure supplement 1 were computed from the actions and rewards of the covariance model by assuming the Q-learning model (Eqs. 1 and 2).

The motor cortex recordings (Fig. 2 – Figure supplement 2)

The data in Fig. 2 – Figure supplement 2A-B was recorded from one female monkey (Macaca fascicularis) at 3 years of age, using a 10x10 microelectrode array (Blackrock Microsystems) with 0.4mm inter-electrode distance. The array was implanted in the arm area of M1, under anesthesia and aseptic conditions.
Behavioral Task: The Monkey sat in a behavioral setup, awake and performing a Brain Machine Interface (BMI) and sensorimotor combined task. Spikes and Local Field Potentials were extracted from the raw signals of 96 electrodes. The BMI was provided through real-time communication between the data acquisition system and a custom-made software, which obtained the neural data, analyzed it and provided the monkey with the desired visual and auditory feedback, as well as the food reward. Each trial began with a visual cue, instructing the monkey to make a small hand movement to express alertness. The monkey was conditioned to enhance the power of beta band frequencies (20-30Hz) extracted from the LFP signal of 2 electrodes, receiving a visual feedback from the BMI algorithm. When a required threshold was reached, the monkey received one of 2 visual cues and following a delay period, had to report which of the cues it saw by pressing one of two buttons. Food reward and auditory feedback were delivered based on correctness of report. The duration of a trial was on average 14.2s. The inter-trial-interval was 3s following a correct trial and 5s after error trials. The data used in this paper, consists of spiking activity of 89 neurons recorded during the last second of inter-trial-intervals, taken from 600 consecutive trials in one recording session. Pairwise correlations were comparable to previously reported (Cohen & Kohn, 2011), $r_{SC} = 0.047 \pm 0.17$ (SD), ($r_{SC} = 0.037 \pm 0.21$ for pairs of neurons recorded from the same electrode).

Animal care and surgical procedures complied with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and with guidelines defined by the Institutional Committee for Animal Care and Use at the Hebrew University.

*The auditory cortex recordings* (Fig. 2 – Figure supplement 2)

The auditory cortex recordings appearing in Fig. 2 – Figure supplement 2C-D are described in detail in (Hershenhoren et al., 2014). In short, membrane potential was recorded intracellularly...
from 39 neurons in the auditory cortex of anesthetized rats. 125 experimental sessions were considered. Each session consisted of 370 50 msec tone bursts, presented every 300-1000 msec. For each session, all trials were either 300 msec or 500 msec long. Trial length remained identical throughout a session and depended on smallest interval between two tones in each session. Trials began 50 msec prior to tone burst. For spike detection, data was high pass filtered with a corner frequency of 30Hz. Maximum points that were higher than 60 times the median of the absolute deviation from the median were classified as spikes.

The Basal ganglia recordings (Fig. 3 and Fig. 2 – Figure supplement 3)

The basal ganglia recordings that are analyzed in Figs. 3 and Fig. 2 – Figure supplement 3 are described in detail in (Ito & Doya, 2009). In short, rats performed a combination of a tone discrimination task and a reward-based free-choice task. Extracellular voltage was recorded in the behaving rats from the NAc and VP using an electrode bundle. Spike sorting was done using principal component analysis. In total, 148 NAc and 66 VP neurons across 52 sessions were used for analyses (In 18 of the 70 behavioral sessions there were no neural recordings).

Estimation of action-values from model choices and rewards

To imitate experimental procedures, we regressed the spike count on estimates of the action-values, rather than the subjective action-values that underlay model behavior (to which the experimentalist has no direct access). For that goal, for each session, we assumed that $Q_l(1) = 0.5$ and found the set of parameters $\hat{\alpha}$ and $\hat{\beta}$ that yielded the estimated action-values that best fit the sequences of actions in each experiment by maximizing the likelihood of the sequence. Action-values were estimated from Eq. (1), using these estimated parameters and the sequence of actions and rewards. Overall, the estimated values of the parameters $\alpha$ and $\beta$ were comparable to
the actual values used: on average, $\hat{\alpha} = 0.12 \pm 0.09$ (standard deviation) and $\hat{\beta} = 2.6 \pm 0.7$
(compare with $\alpha=0.1$ and $\beta=2.5$).

Exclusion of neurons

Following standard procedures, a sequence of spike-counts, either simulated or experimentally measured was excluded due to low firing rate if the mean spike count in all blocks was smaller than 1. This procedure excluded 0.02% (4/20,000) of the random-walk neurons. Considering the auditory cortex recordings, we assigned each of the 125 spike counts to 40 randomly-selected sessions. 23% of the neural recordings (29/125) were excluded in all 40 sessions. Because blocks are defined differently in different sessions, some neural recordings were excluded only when assigned to some sessions but not others. Of the remaining 96 recordings, 14% of the recordings × sessions were also excluded. Similarly, considering the basal ganglia neurons, we assigned each of the 642 recordings (214×3 epochs) to 40 randomly-selected sessions. 11% (74/(214×3)) of the recordings were excluded in all 40 sessions. Of the remaining 568 recordings, 9% of the recordings×sessions were also excluded. None of the simulated action-value neurons (0/20,000) or the motor cortex neurons (0/89) were excluded.

Statistical analyses

The computation of the t-values of the regression of the spike counts on the estimated Q-values (as in Figs. 1, 2, 5, Fig. 2 – Figure supplement 1, – Figure supplement 2, – Figure supplement 3) was done using the following regression model:

$$s(t) = \beta_0 + \beta_1 Q_1(t) + \beta_2 Q_2(t) + \epsilon(t) \quad (6)$$

Where $s(t)$ is the spike count in trial $t$, $Q_1(t)$ and $Q_2(t)$ are the estimated action-values in trial $t$, $\epsilon(t)$ is the residual error in trial $t$ and $\beta_0$–$\beta_2$ are the regression parameters.
The computation of the t-values of the regression of the spike counts on the reward probabilities in the trial design experiment (as in Fig. 4) was done using the following regression model:

\[ s(t) = \beta_0 + \beta_1 RP_1(t) + \beta_2 RP_2(t) + \epsilon(t) \] (7)

Where \( s(t) \) is the mean spike count in the last 200 trials of the session, \( RP_1(t) \) and \( RP_2(t) \) are the reward probabilities given action 1 or action 2 were chosen, respectively in the last 200 trials of the session (in this experimental design \( RP \) could be 0.1, 0.5 or 0.9), \( \epsilon(t) \) is the residual error and \( \beta_{0-2} \) are the regression parameters.

The computation of the t-values of the regression of the spike counts on state and policy in a trial design experiment was done using the following regression model:

\[ s(t) = \beta_0 + \beta_1 [RP_1(t) + RP_2(t)] + \beta_2 [RP_1(t) - RP_2(t)] + \epsilon(t) \] (8)

All variables and parameters are the same as in Eq. (7)

All regression analyses were done using regstats in MATLAB (version 2016A).

To find neurons whose spike count in the last 20 trials is modulated by reward probability (Figs. 1B, 2A) we executed the Wilcoxon rank sum test, using ranksum in MATLAB. All tests were two-tailed.

Significance of t-values slightly depends on session length. For the session lengths we considered, 0.05 significance bounds varied between 1.962 and 1.991. For consistency, we chose a single conservative bound of 2. Similarly, 0.025 and 0.01 significance bounds were chosen to be 2.3 and 2.64, respectively.

For all significance boundaries the false positive thresholds were computed naively, i.e., assuming the analysis is not confounded in any way and that the two predictors are not correlated with each other. For example, assuming the false positive rate from a single t-test for a
significant regression coefficient is $P$, for the standard analysis, the false positive rate for each action-value classification was defined as $P \cdot (1 - P)$, and the false positive rate was equal for state and policy classification and was defined as $P^2/2$. In Fig. 6 the false positive rate computed for random-walk neurons was $P^2/2$ for each action-value, and the false positive rate computed for state or policy neurons was $P/2$ for each action-value.

Permutation test (Fig. 3)

For each neuron, we computed the t-values of the regressions of its spike-count on estimated action-values from the sessions of Fig. 1E. Because the number of trials can affect the distribution of t-values, we only considered in our analysis the first 170 trials of the 504 sessions longer or equal to 170 trials. This number, which is approximately the median of the distribution of number of trials per session, was chosen as a compromise between the number of trials per session and number of sessions.

Two points are noteworthy. First, the distribution of the t-values of the regression of the spike count of a neuron on all action-values depends on the neuron (see Fig. 3A). Similarly, the distribution of the t-values of the regression of the spike counts of all neurons on an action-value depends on the action-value (not shown). Therefore, the analysis could be biased in favor (or against) finding action-value neurons if the number of neurons analyzed from each session (and therefore are associated with the same action-values) differs between sessions. Second, this analysis does not address the correlated decision variables confound.

Finally, we would like to point out that there is an alternative way of performing the permutation test, which is applicable when the number of sessions is small, while the number of neurons recorded in a session is large. Instead of comparing the t-values from the regression of a neuron
on different action-values, one can compare the t-values from different neurons on the same action-value. However, this method is only applicable under the assumption that the temporal correlations that are not related to action-value in the neuronal activity are similar between sessions.

Comparison with permuted spike counts (Fig. 2–Figure supplement 4)

In Fig. 2–Figure supplement 4 we considered the experiment and analysis described in (H. Kim et al., 2009). That experiment consisted of four blocks, each associated with a different pair of reward probabilities, (0.72, 0.12), (0.12, 0.72), (0.21, 0.63) and (0.63, 0.21), appearing in a random order, with the better option changing location with each block change. The number of trials in a block was preset, ranging between 35 and 45 with a mean of 40 (this is unlike the experiment described in Fig. 1, in which termination of a block depended on performance).

First, we used Eqs. (1) and (2) to model learning behavior in this protocol. Then, we estimated the action-values according to choice and reward sequences, as in Fig. 1. These estimated action-values were used for regression of the spike counts of the random-walk, motor cortex, auditory cortex, and basal ganglia neurons in the following way: each spike count sequence was randomly assigned to a particular pair of estimated action-values from one session. The spike count sequence was regressed on these estimated action-values. The resultant t-values were compared with the t-values of 1,000 regressions of the spike-count, permuted within each block, on the same action-values. The p-value of this analysis was computed as the percentage of t-values from the permuted spike-counts that were higher in absolute value than the t-value from the regression of the original spike count. The significance boundary was set at p<0.025 (H. Kim et al., 2009). Neurons with at least one significant regression coefficient (rather than exactly one
significant regression coefficient) were classified as action-value modulated neurons (H. Kim et al., 2009).

ANOVA tests for comparisons between blocks, excluding "drifting" neurons

Following (Asaad et al., 2000) we conducted an additional analysis with repeating blocks. We simulated learning behavior in the same experiment as in Fig. 2 – Figure supplement 10. This experiment is composed of 8 blocks - the 4 blocks of Fig. 1, repeated twice, in random permutation. We restricted our analysis to the 438 sessions with 332 trials or fewer (332 trials is the shortest session in the basal ganglia recording). Each spike count was analyzed 40 times, using 40 randomly-assigned sessions. For each block, we restricted the analysis to the neuronal activity in the last 20 trials of the block.

First, we conducted four one-way ANOVAs (using MATLAB’s anova1) to compare the neuronal activities in blocks associated with the same action-values (e.g., the neuronal activity in the two blocks, in which reward probabilities were (0.1,0.5)). Neurons were excluded from further analysis if we found a significant difference in their firing rates in at least one of these comparisons (df(columns)=1, df(error)=38 , p<0.1). This procedure excludes from further analysis "drifting" neurons, whose spike count significantly varied in the session.

Next, for each action-value we conducted a one-way ANOVA (using MATLAB’s anova1), which compared the neuronal activity between the two blocks in which the action-value was 0.1 and the two blocks in which the action-value was 0.9 (df(columns)=1, df(error)=78 , p<0.01). We classified neurons as representing action-values if there was a significant difference between their firing rates for one action-value but not for the other.
Despite the removal of "drifting" neurons, this analysis yielded an erroneous classification of action-value neurons in all datasets: random-walk neurons, 18%; motor cortex neurons, 12%; auditory cortex neurons, 5%; basal ganglia neurons, 9%. This is despite the fact that the expected false positive rate is only 2%. These results indicate that the exclusion of "drifting" neurons as in (Asaad et al., 2000) does not solve the temporal correlations confound.

Data from the motor cortex, auditory cortex, and basal ganglia was the same as in Fig. 2 – Fig. supplements 2-3. Data for random-walk included 1000 newly simulated neurons, using the same parameters as in Fig. 2 (this was done to create enough trials in each simulated spike count).

Data and Code Availability

The data of the basal ganglia recordings is available online at https://groups.oist.jp/ncu/data and was analyzed with permission from the authors. Motor cortex and auditory cortex data is available at https://github.com/lotem-elber/striatal-action-value-neurons-reconsidered-codes (Elber-Dorozko 2018). The custom MATLAB scripts used to create simulated neurons and to analyze simulated and recorded neurons are also available at https://github.com/lotem-elber/striatal-action-value-neurons-reconsidered-codes.

Acknowledgements

We are extremely grateful to Oren Peles, Eilon Vaadia and Uri Werner-Reiss for providing us with their motor cortex recordings, Bshara Awwad, Itai Hershenhoren, Israel Nelken for providing us with their auditory cortex recordings, Kenji Doya and Makoto Ito for providing us with their basal ganglia recordings, Mati Joshua, Gianluigi Mongillo, Jonathan Roiser and Roey Schurr for careful reading of the manuscript and helpful comments and Inbal Goshen, Hanan Shteingart and Wolfram Schultz for discussions.
References


comparisons in the dorsolateral prefrontal cortex control choice between goal-directed actions. Nature Communications, 5, 1–10. https://doi.org/10.1038/ncomms5390


Pasquereau, B., Nadjar, A., Arkadir, D., Bezard, E., Goillandeau, M., Bioulac, B., Gross, C. E.,
https://doi.org/10.1523/JNEUROSCI.3745-06.2007


https://doi.org/10.1126/science.1115270

*Physiological Reviews*, 95, 853–951. https://doi.org/10.1152/physrev.00023.2014

https://doi.org/10.1016/j.neuron.2012.03.037

6273(03)00761-X


correlates of stimulus – response and response – outcome associations in dorsolateral versus
dorsomedial striatum. *Frontiers in Integrative Neuroscience, 4*, 1–18.

https://doi.org/10.3389/fnint.2010.00012


Erroneous detection of action-value representation in a model with covariance based synaptic plasticity.

(A) An example of operant learning of the covariance model (see Materials and Methods). Legend is the same as in Fig 1A. (B) Two example covariance neurons that appear if they represent action-values. The red (top) and blue (bottom) lines denote the calculated action-values 1 and 2, respectively that were computed from the behavior of the model. Gray lines and gray dots denote the firing rates and the spike counts of two example covariance neurons. Black horizontal lines denote the mean spike count in the last 20 trials of the block. Error bars denote the standard error of the mean. The two asterisks denote p<0.01 (rank sum test). Legend is the same as in Fig 2A. (C) and (D) Population analysis. Same as in Figs. 1D and 1E. Each simulated neuron was regressed on the computed action-values. The two simulated neurons in (B) are denoted by squares in (C). Results in (D) are based on 500 sessions with 2,000 simulated neurons in a session. Legend is the same as in Fig. 1D and 1E, respectively. Dashed lines in (C) at t=2 denote the significance boundaries. Dashed lines in (D) denote the naïve expected false positive rate from the significance threshold (see Materials and Methods). Error bars denote standard error of the mean. Following this standard approach, 43% of the covariance neurons would have been erroneously classified as representing action-values.
Erroneous detection of action-value neurons in unrelated experiments. (A) and (B) analysis on motor cortex neurons. (A) An example motor cortex neuron recorded in a BMI task, presented as if the sequence of spike counts of this neuron corresponds to the sequence of trials in a randomly chosen session of operant learning from the sessions used for the population analysis in Fig. 1E. Gray dots denote the spike-counts. Black horizontal line denotes the mean spike counts in the last 20 trials of the assigned blocks. Error bars denote the standard error of the mean. The two asterisks denote p<0.01 (rank sum test). For each neuron, we computed the t-values of the regression of the spike count on the two corresponding estimated action-values. The red line denotes the action-value whose t-value exceeded 2 (in absolute value). (B) Population analysis. Left scatter plot, the t-values of 89 neurons regressed on the estimated action-values of randomly selected 89 sessions (same as Fig. 1D). The neuron in (A) is denoted by a square. Dashed lines at t=2 denote the significance boundaries. Right bar chart, fraction of neurons classified in each category, estimated by regressing each of the 89 motor cortex neurons on 80 different estimated action-values from 40 randomly selected sessions. Dashed lines denote the naïve expected false positive rate from the significance threshold (see Materials and Methods). Error bars denote the standard error of the mean. Legend is the same as in Fig. 1D and 1E. (C) and (D) analysis on auditory cortex neurons of (Hershenhoren et al., 2014). (C) Same as in (A) for an auditory cortex neuron in an anesthetized rat responding to the presentation of pure tones. (D) Population analysis. Left scatter plot, the t-values of 82 recorded sessions from auditory neurons regressed on the estimated action-values of randomly selected 82 sessions (same as (B)). The neuron in (C) is denoted by a square. Right bar chart, fraction of neurons classified in each category, estimated by regressing 125 recorded sessions from auditory cortex neurons on 80 different estimated action-values from 40 randomly selected sessions (in each session, 34% of recordings were excluded on average, see Materials and Methods). Error bars
denote the standard error of the mean. Following this standard approach, 36% of the motor cortex neurons and 23% of the auditory cortex neurons were erroneously classified as representing action-values.
Figure 2 – Figure supplement 3  Erroneous detection of unrelated action-values in basal ganglia neurons. Population analysis on basal ganglia neurons. Spike counts were regressed on estimated action-values created in the same experimental setting as in Fig. 1. To compare with the number of blocks and trials used in the original analysis, we simulated sessions with more blocks, so that the original 4 blocks were repeated each time in random permutation. The average number of blocks and trials used in this analysis is 6 and 516.5, respectively. Left scatter plot, the t-values of 214 neurons of (Ito & Doya, 2009) in three different phases regressed on the estimated action-values from randomly selected 642 simulated sessions (same analysis as in Fig. 2 – Figure supplement 2B,D). Dashed lines at t=2 denote the significance boundaries. Right bar chart, fraction of neurons classified in each category, estimated by regressing 214 neurons in three different phases on 80 different estimated action-values from 40 randomly selected sessions (see Materials and Methods). Dashed lines denote the naïve expected false positive rate from the significance threshold (see Materials and Methods). Error bars denote the standard error of the mean. Legend is the same as in Fig. 1D and 1E. This analysis erroneously identified 43% of the neurons as action-value neurons, despite the fact that these action-values were completely unrelated to the experimental session in which these neurons were recorded. These results demonstrate that the magnitude of non-stationarity in standard electrophysiological recordings is sufficient to result in an erroneous identification of neurons as representing action-values.
Figure 2 – Figure supplement 4 Spike count permutation (as in (H. Kim et al., 2009)) does not resolve the temporal correlations confound. Conducting this analysis using four data sets and unrelated, simulated action-values we erroneously classify action-value representation in all data sets. (A), (B) (C) and (D) denote the random-walk neurons, motor cortex neurons, auditory cortex neurons and basal ganglia neurons, respectively. Left, t-values from regressions of the original spike-count on the estimated action-values. Green triangles denote significant modulation by action-value according to the permuted spike-count analysis (see Materials and Methods). Dashed black lines at t=2 denote the significance boundaries that would be used in the standard analysis. Right, fraction of neurons significantly modulated by action-value according to the permuted spike-count analysis across the population (0.05, expected by chance, is marked by a horizontal dashed line). Error bars are standard error of the mean. Number of neurons used in (A), (B) (C) and (D) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-B, Fig. 2 – Figure supplement 2C-D and Fig. 2 – Figure supplement 3, respectively. Legend is the same as in Fig. 1D. Note that in all four cases the two t-values are correlated. This results from the correlation between $Q_1(t)$ and $Q_2(t)$ caused by the reward schedule in (H. Kim et al., 2009).
Following (H. Kim et al., 2013), we simulated the experimental settings as in (H. Kim et al., 2013) to create simulated action-values (see also Materials and Methods on Fig. 2 – Figure supplement 4, with the same experimental design, but with a different statistical analysis) and conducted a multiple linear regression analysis on each of the unrelated data sets, using the simulated action-values and the following regression model:

\[
s(t) = \beta_0 + \beta_1 Q_1(t) + \beta_2 Q_2(t) + \beta_3 C(t) + \beta_4 R(t) + \beta_5 C(t) \cdot R(t) + \beta_6 CV(t) + \beta_7 s(t-1) + \beta_8 s(t-2) + \beta_9 s(t-3) + \epsilon(t)
\]

where \(s(t)\) is the spike count in trial \(t\), \(Q_1(t)\) and \(Q_2(t)\) are the estimated action-values in trial \(t\), \(C(t)\) is the action chosen in trial \(t\), \(R(t)\) is the reward in trial \(t\), \(C(t) \cdot R(t)\) is the interaction between choice and reward in trial \(t\), where both are expressed as binary with values \{-1,1\}, \(CV(t)\) is the value of the action that was chosen on trial \(t\), \(s(t-1)\), \(s(t-2)\), \(s(t-3)\) are the spike counts one, two and three trials prior to current trial, respectively, \(\epsilon(t)\) is the residual error in trial \(t\) and \(\beta_{0-9}\) are the regression parameters. \(\text{(A)}\), \(\text{(B)}\) \((\text{C})\) and \(\text{(D)}\) denote the random-walk neurons, motor cortex neurons, auditory cortex neurons and basal ganglia neurons, respectively. Left, \(t\)-values from regressions of the spike-count on the regression variables. The significance boundaries for the \(t\)-values, denoted by dashed lines, are 2.3, corresponding to \(p<0.025\). Right, fraction of neurons significantly modulated by action-value across the population (0.05, expected by chance, is marked by a horizontal dashed line). Error bars are standard error of the mean. Number of neurons used in \(\text{(A)}\), \(\text{(B)}\) \((\text{C})\) and \(\text{(D)}\) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-
B, Fig. 2 – Figure supplement 2C-D and Fig. 2 – Figure supplement 3, respectively. Legend is the same as in Fig. 1D. Note that in all four cases the two t-values are correlated. This results from the correlation between $Q_1(t)$ and $Q_2(t)$ caused by the reward schedule in (H. Kim et al., 2013).
Figure 2 – Figure supplement 6 Regression on reward probabilities does not resolve the temporal correlations confound. Following (Ito & Doya, 2009; Samejima et al., 2005) for each of the four data-sets (A) random-walk (B) motor cortex (C) auditory cortex and (D) basal ganglia neurons, spike counts in the last 20 trials in each block (from randomly assigned simulated experimental settings, the assignment was the same as in the standard analysis in Figs. 2 and 2–figure supplements 2,3) were regressed on reward probabilities (e.g., (0.5,0.9)) in those blocks. This is similar to the analysis in the individual examples of Figs. 1B, 2A, Fig. 2 – figure supplement 2A, and Fig. 2 – figure supplement 2C (in which two rank sum tests, and not regression, were used). Left of each panel denotes the t-values of the regressions of individual neurons (Dashed lines at t=2 denote the significance boundaries) and right bar graphs denote the population statistics (Dashed lines denote the naïve expected false positive rate from the significance threshold, see Materials and Methods). Number of neurons used in (A), (B) and (C) and (D) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-B, Fig. 2 – Figure supplement 2C-D and Fig. 2 – Figure supplement 3, respectively. Legend is the same as in Fig. 1D and 1E. Note that for this analysis we considered significance using the threshold of p<0.05. By contrast, in (Ito & Doya, 2009) the same analysis was used with a significance threshold of p<0.01. For comparison, when considering the basal ganglia neurons with a significance threshold of p<0.01, the number of neurons that are erroneously classified as action-value neurons decreases from 37%±3.3% to 26%±3%.
Figure 2 – Figure supplement 7 Detrending analysis does not resolve the temporal correlations confound. Following (Ito & Doya, 2015a), we conducted a multiple linear regression analysis using unrelated action-values (the same action-values as in Fig. 2 – figure supplement 6) and the following regression model:

\[ s(t) = \beta_0 + \beta_1 Q_1(t) + \beta_2 Q_2(t) + \beta_3 t + \beta_4 C(t) + \beta_5 C(t-1) + \beta_6 R(t) + \beta_7 R(t-1) + \epsilon(t) \]

Where \( s(t) \) is the spike count in trial \( t \), \( Q_1(t) \) and \( Q_2(t) \) are the estimated action-values in trial \( t \), \( C(t) \) and \( C(t-1) \) are the actions chosen in trial \( t \) and \( t-1 \), respectively, \( R(t) \) and \( R(t-1) \) are the rewards in trial \( t \) and \( t-1 \), respectively, \( \epsilon(t) \) is the residual error in trial \( t \) and \( \beta_{0-7} \) are the regression parameters. (A), (B) (C) and (D) denote the random-walk neurons, motor cortex neurons, auditory cortex neurons and basal ganglia neurons, respectively. Left, t-values from regressions of the spike-count on the regression variables. As in (Ito & Doya, 2015a), the significance boundaries for the t-values, denoted by dashed black lines, are 2.64, corresponding to \( p<0.01 \) (as opposed to \( p<0.05 \) elsewhere). Right bar graphs denote the population statistics. Dashed lines denote the naive expected false positive rate from the significance threshold (see Materials and Methods). Note, however, that the significance criterion is more stringent and the expected total number of identified action-value neurons by chance is only 2%. Number of neurons used in (A), (B) (C) and (D) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-B, Fig. 2 – Figure supplement 2C-D and Fig. 2 – Figure supplement 3, respectively. Legend is the same as in Fig. 1D and 1E.
Unbiased identification of action-value neurons does not resolve the temporal correlations confound. Following (Wang et al., 2013) (whose main focus was state-value representation), we considered an unbiased identification of action-value neurons. (A), (B) (C) and (D) denote the random-walk neurons, motor cortex neurons, auditory cortex neurons and basal ganglia neurons, respectively. The t-values for the different neurons are identical to Fig. 2 – Figure supplement 6 (and unlike (Wang et al., 2013) this analysis used only the last 20 trials in each block). The f-value of each neuron was computed from the regression and a neuron was considered as non-significant (black dot) if p > 0.01, denoted by the circle in the left panels. For the significant neurons, the dashed lines define 8 equal-angle sectors, each corresponding to a different classification of the neuron. Right is the population analysis. Dashed lines denote the naïve expected false positive rate from the significance threshold (see Materials and Methods). Note that the expected total number of identified significant neurons by chance is only 1%. Number of neurons used in (A), (B) (C) and (D) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-B, Fig. 2 – Figure supplement 2C-D and Fig. 2 – Figure supplement 3, respectively. Legend is the same as in Fig. 1D and 1E.
Figure 2 – Figure supplement 9 Random intermingling of estimated action-values does not resolve the temporal correlations confound. In the spirit of the experiment conducted by (Fitzgerald et al., 2012) we simulated an experiment, in which two different trial types are marked by cues, randomly selected every trial, and action-values are learned separately for each cue. Specifically, each session in this new design was created by a random intermingling of two different, randomly-selected sessions from those analyzed in Fig. 1. The number of trials of the intermingled session was equal to that of one of the two randomly-selected original sessions (i.e., we only used approximately the first half of each of the original sessions). We created 1,000 such intermingled sessions. Next, we regressed the spike counts of neurons from each of the four data-sets (A) random-walk neurons, (B) motor cortex neurons and (C) auditory cortex neurons and (D) basal ganglia neurons on the resulting intermingled estimated action-values. Left of each panel denotes the t-values of the regressions of individual neurons (Dashed lines at t=2 denote the significance boundaries) and right bar graphs denote the population statistics (Dashed lines denote the naïve expected false positive rate from the significance threshold, see Materials and Methods). Number of neurons used in (A), (B), (C) and (D) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-B, Fig. 2 – Figure supplement 2C-D and Fig. 2 – Figure supplement 3, respectively. Legend is the same as in Fig. 1D and 1E. These results show that even when using a design where trials are chosen randomly, there can still be temporal correlations in the predictors of the model. In this case, this occurs because the temporal correlations in each estimated action-value still create temporal correlations in the intermingled vector.
Figure 2 – Figure supplement 10 Increasing the number of blocks does not resolve the temporal correlations confound. Standard analysis with 8 blocks. The 4 blocks from the experiment in Fig. 1A were repeated twice, each time in random permutation. The mean length of the sessions was 347 trials (standard deviation 65 trials). For each of the three data-sets (A) random-walk neurons, (B) motor cortex neurons and (C) auditory cortex neurons longer sessions with 8 blocks were simulated and then spike-counts were regressed on the longer estimated action-values (for auditory cortex analysis only 676 sessions with 370 or fewer trials were used). Left of each panel denotes the t-values of the regressions of individual neurons (Dashed lines at t=2 denote the significance boundaries) and right bar graphs denote the population statistics (Dashed lines denote the naïve expected false positive rate from the significance threshold, see Materials and Methods). Number of neurons used in (A), (B) and (C) is the same as in Figs. 2, Fig. 2 – Figure supplement 2A-B and Fig. 2 – Figure supplement 2C-D, respectively. Legend is the same as in Fig. 1D and 1E. We did not perform this analysis on the basal ganglia neurons because we already used longer sessions in the original analysis for these recordings (see Fig. 2 – figure supplement 3).
Figure 3 – Figure supplement 1 (A), the standard analysis on the neuronal recordings taken from (Ito & Doya, 2009), using action-values estimated from the behavior in the sessions in which the neurons were recorded. The top and bottom scatter plots are identical, except for the significance boundries. They depict the t-values from the results of the regression of 640 of 642 neuronal recordings (214 neurons × 3 epochs) on the action-values that were estimated from the original experiment (see Materials and Methods for estimation of action-values). Legend is the same as in Fig. 1D and 1E (Dashed lines in the right panels at t=2 denote the significance boundaries and the dashed lines on the left panels denote the naïve expected false positive rate from the significance threshold, see Materials and Methods). This analysis is different from the one used in from (Ito & Doya, 2009), which is similar to the one reported in (Fig. 2 – figure supplement 6). Two neurons do not appear on the scatter plots, whose axes were bounded for ease of viewing. Their t-values were (2.44, 18.91) and (1.08, 16.96) for action-value 1 and 2, respectively. (B) For comparison, we repeated the analysis for the random-walk neurons (Fig. 2). The fraction of erroenusly-identified action-value neurons is comparable to that extracted from the experimental data. Remarkably, bias in favor of 'state'-representing neurons over 'policy'-representing neurons observed in the recorded neurons is also present in the random-walk neurons. This suggests that the overrepresentation of state neurons is not necessarily of biological significance.
Figure 5 – Figure supplement 1 Alternative analyses on policy neurons. (A) Regression analysis on policy neurons with choice as an added regressor. Following common experimental practice, we used the following regression model: $s(t) = \beta_0 + \beta_1 Q_1(t) + \beta_2 Q_2(t) + \beta_3 (a(t) = 1) + \epsilon(t)$ where $s(t)$ is the spike count in trial $t$, $Q_1(t)$ and $Q_2(t)$ are the estimated action-values in trial $t$, $(a(t) = 1)$ is a binary variable indicating whether action 1 was chosen, $\epsilon(t)$ is the residual error in trial $t$ and $\beta_0 \ldots \beta_3$ are the regression parameters. Simulated neurons are the same as in Fig. 5E. Legend is the same as Figs. 1D and 1E (Dashed lines in the right panels at $t=2$ denote the significance boundaries and the dashed lines on the left panels denote the naïve expected false positive rate from the significance threshold, see Materials and Methods). (B) Unbiased identification of action-value representation in policy neurons. Following (Wang et al., 2013), we considered an unbiased identification of action-value neurons. The $f$-value of each neuron was computed from the regression and a neuron was considered as non-significant (black dot) if $p>0.05$, denoted by the circle in the left figure. For the significant neurons, the dashed lines define 8 equal-angle sectors, each corresponding to a different classification of the neuron, similarly to Fig. 2–Figure supplement 8. The figure on the right is the population analysis (Dashed lines denote the naïve expected false positive rate from the significance threshold, see Materials and Methods). Note that the expected number of neurons that will be classified as action-value neurons by chance is only 2.5%. Simulated neurons are the same as in Fig. 5E. Legend is the same as in Fig. 1D and 1E. In the original paper, $p<0.01$ was used. For $p<0.01$ the analysis classifies 6% of neurons as action-value neurons (0.5% expected by chance) and 17% as $\Delta Q$ neurons (0.25% expected by chance).