Striatal action-value neurons reconsidered

Lotem Elber-Dorozko[1] and Yonatan Loewenstein[1,2]

[1]The Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem.

[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 a large number of electrophysiological studies, in which the neural activity of striatal neurons was measured while the subject was learning to prefer the more rewarding alternative. Here we present an alternative interpretation of the electrophysiological findings. We show that the standard statistical methods that were used to identify action-value neurons in the striatum erroneously detect the same action-value representations in unrelated neuronal recordings. This is due to temporal correlations in the neuronal data. We propose an alternative statistical method for identifying action-value representations that is not subject to this caveat. We apply it to previously identified action-value neurons in the basal ganglia and fail to detect action-value representations. In conclusion, we argue that there is no conclusive evidence for the generally accepted hypothesis that striatal neurons encode action-values.

Key words

action-value; striatum; temporal correlations;

Introduction

There is a long history of operant learning experiments, in which a subject, human or animal, repeatedly chooses between actions and is rewarded, often stochastically, according to its choices. A popular theory posits that the subject's decisions in this task 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[1]. In recent years, there is a lot of interest in the neural mechanisms underlying this computation[2,3]. In particular, based on several electrophysiological 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.

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: in a multi-armed bandit task with binary rewards, the average reward associated with an action can be defined as the corresponding probability of reward. 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[4,5]. 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:

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

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

In the framework of a two-alternative task with binary rewards, $i \in \{1,2\}$, $a(t) \in \{1,2\}$ and $R(t) \in \{1,0\}$ are the possible actions, choice and reward in trial t, respectively, α is the learning rate and Q_i is the action-value associated with action i.

It is typically assumed that the probability of choosing an action is a softmax sigmoidal function of the difference of the action-values (see also[6]):

$$\Pr(a(t) = 1) = \frac{1}{1+e^{-\beta(Q_1(t)-Q_2(t))}} \qquad (2)$$

where β is a parameter that determines the tradeoff between exploration and exploitation, the bias towards the action associated with the higher action-value.

The parameters of the model, α and β, can be estimated from the behavior, allowing the researchers to compute Q_1 and Q_2 on a trial-by-trial basis.

By measuring neural activity while the subject is performing the operant task, computing the regression of the trial-by-trial spike counts of the neurons on the latent variables $Q_i(t)$ and identifying neurons for which this regression is statistically significant, one can identify the neurons that represent action-values.

Using this framework, several electrophysiological studies in the past decade have found that the firing rate of a substantial fraction of striatal neurons (12%-40% for different significance thresholds) is significantly correlated with the average reward associated with one of the actions, regardless of whether the action was chosen. These results were considered as evidence that neurons in the striatum represent action-values[4,5,7,8].

In this paper we point out a potential caveat about this analysis - it could result in erroneous identification of neurons as representing action-value, if the firing rates are temporally correlated. The paper is organized in the following way. We commence by describing a standard method for identifying action-value neurons. Then, we use numerical simulations to show in a model that this method also erroneously identifies action-value neurons when they do not exist. Next, we apply this method to neuronal recordings from the motor cortex of a monkey performing an unrelated task and to neuronal recordings from the auditory cortex of anaesthetized rats and demonstrate that this method also erroneously identifies action-value neurons in these recordings. Finally, we demonstrate erroneous identification of action-value neurons in the basal ganglia of behaving rats. In the Supplementary Information section, we show that the previously-proposed alternative methods for identifying action-value neurons suffer from the same caveat. We conclude by proposing a different analysis for identifying action-value neurons, that is not subject to this caveat and applying it to basal ganglia recordings, in which action-value neurons were previously identified. Using this new analysis we fail to detect action-value representations.

Results

Identifying action-value neurons

We commence by examining the standard methods of identifying action-value neurons using a simulation of an operant learning experiment. We considered 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[4,8].

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[5]) and initial conditions $Q_i(1) = 0.5$). 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. 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 firing rates after learning by comparing the spike counts 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 Q-values, estimated from the subject's history of choices and rewards[4,5,7–10]. 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).

Next, we regressed the spike count of each simulated neuron on the two estimated action-values corresponding to its 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 ($t_{18}(Q_1) = 4.05$) but not for the estimated Q_2 ($t_{18}(Q_2) = -0.27$). Similarly, the t-values of the regression coefficients of the blue-labeled action 2-value neuron was significant for the estimated Q_2 ($t_{18}(Q_2) = 3.05$) but not for the estimated Q_1 ($t_{18}(Q_1) = 0.78$).

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 in the simulation were identified as action-value neurons. Only 2% of the simulated neurons had significant regression coefficients

with both action-values. These neurons are typically classified as state or policy neurons, if the two regression coefficients have the same or different signs, respectively[8]. Note that, for two reasons, not all simulated neurons were identified as action-value neurons: first, the estimated action-values are not identical to the underlying Q-values, which determine the firing rates 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.

Identifying "action-value" neurons in the absence of value (model)

The identification of the simulated neurons in Fig. 1d,e as action-value neurons relied on the interpretation that a large t-value is highly improbable under the null hypothesis that the firing rate of the neuron is not modulated by action-values. However, when computing the significance threshold for rejection of the null hypothesis it was implicitly assumed that spike counts in different trials are independent. 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 rate follows 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 to the trials in the simulation depicted in Fig. 1a, and

considered the spike counts of the random-walk neurons in the last 20 trials of each of the four blocks in Fig. 1a. 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), similar 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 Q-values from Fig. 1e and computed the t-values (Fig. 2b). This analysis classifies 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 action 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 using this method as a function of the diffusion parameter σ (Fig. 2d). When $\sigma=0$, the firing rates 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%, as expected from a significance criterion of 5% and two statistical tests, corresponding to the two action-values. The larger the value of σ, 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.

The excess action-value neurons in Fig. 2 emerged because the statistical analysis was based on the assumption that the spike counts in different trials are independent from each other. This independence assumption is violated in the case of a random-walk process, in which the firing rate

randomly drifts with time, leading to correlations between spike counts in temporally proximate trials. Because of this non-stationarity, the expected variability in the statistic (be it average spike count in 20 trials or the regression coefficient), which is calculated under the independence assumption, is an underestimate of the actual variability. Therefore, the fraction of random-walk neurons classified as action-value neurons increases with the magnitude of the drift, which is directly related to the magnitude of correlations between spike counts in proximate trials (Fig. 2d).

Importantly, the Gaussian random-walk process is just one example in which the firing rate is non-stationary. Other processes, in which the firing rate is non-stationary (e.g., oscillatory or trend following) and thus the independence assumption is violated may also lead to an erroneous identification of neurons as action-value neurons. For example, it has been suggested that synaptic plasticity that stochastically implements or approximates direct policy gradient learning underlies some forms of operant learning[11–16]. In general, there will be no explicit or implicit representation of action-values when these algorithms are implemented. However, because neural activity in these algorithms slowly varies over trials, the methods described above for identifying action-value neurons may erroneously identify action-value representations in implementations of these algorithms. To test this, we studied learning mediated by covariance-based synaptic-plasticity[15,17–19] in the learning task of Fig. 1a (Supplementary Information). Indeed, not only did this algorithm successfully learn to prefer the better alternative, when considering the spike counts of the simulated neurons in this algorithm, 43% of these neurons were erroneously identified as action-value neurons (Fig. S1).

Identifying "action-value" neurons in the absence of value (experiments)

In the previous section we demonstrated that the standard analysis depicted in Fig. 1 may lead to the erroneous identification of neurons as action-value neurons if the firing rate is sufficiently non-

stationary. To test whether this theoretical finding is relevant to electrophysiological experiments, we considered the spike count of 89 single neurons recorded extracellularly from the motor cortex of a monkey engaged in a brain-machine-interface (BMI) experiment. The task was composed of 600 identical trials, in which the monkey was conditioned to increase the amplitude of beta oscillations in the motor cortex and afterwards rewarded for reporting correctly on a viewed visual cue (Materials and Methods). For the purpose of our analysis, we considered as a "trial" the spike count of the neuron in the last 1 sec of each inter-trial-interval in the original experiment.

As in the analyses in Fig. 2, every spike count sequence of a motor cortex neuron was randomly paired with a pair of estimated Q-values from one of the simulations of the operant task depicted in Fig. 1 (truncating the number of experimentally-measured trials in accordance with the number of trials in the simulation). Fig. 3a depicts two estimated Q-values from two sessions (lines), imposed on the spike counts (dots) of two motor cortex neurons. Similar to the random-walk neurons (Fig. 2a), we compared the spike count in the last 20 trials of each of the four blocks and found that the sequence of spike counts of the Top neuron in Fig. 3a matched that of an action 1-value neuron. The sequence of spike counts of the Bottom neuron in Fig.3a matched that of an action 2-value neuron.

For the population analysis, we regressed all vectors of motor-cortex spike counts on the estimated Q-values of Fig. 1. Similarly to the results of the simulations of the random-walk neurons, 36% of the motor cortex neurons in this experiment were classified as action-value neurons (fig. 3b,c). 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.

To test whether this limitation of the analysis is restricted to extracellular recordings, we considered intracellular recordings of 39 auditory cortex neurons in 125 sessions (of which 29 sessions were excluded in all repetitions of the analysis due to low spike count, see Materials and Methods) in anaesthetized rats, responding to auditory stimuli[20] (Materials and Methods). In short, the animals were exposed to a long sequence of pure tones, presented every 300-1000 msec. Depending on the session, trials in our analysis were taken to be 300 msec or 500 msec long (trial length remained the same throughout a session) and included a single pure tone. Repeating the same analysis on these auditory cortex neurons (Fig. 4), 23% of the neurons passed the selection criterion for action-value neurons (Fig. 4b-c; see individual examples in Fig. 4a).

Identifying representations of unrelated "action-value" in the basal ganglia

To test whether the erroneous identification of action-value neurons in the motor and auditory cortices is relevant to the statistics of firing in the striatum, we considered recordings from the nucleus accumbens (NAc) and ventral pallidum (VP) of rats in an operant learning experiment[21].The experiment was a combination of a tone discrimination task and a reward-based free-choice task. We considered only the free choice trials, in which the appropriate response of the animal was to perform a nose poke in either the left or right hole after exiting the center hole. The experimental session was composed of 4-11 blocks. The blocks were the same as the ones presented in Fig. 1a, randomly ordered and (in most cases) repeatedly presented. Blocks changed when the subject chose the higher-valued action at least 80% of the time in 20 trials. As in[21], we considered the spike count in three 1 sec phases of the trial - before nose poking in the central hole, 1 sec following initiation of choice-instruction tone and last 1 sec of nose poke in central hole. In what follows we aggregate the three phases.

For each recording, we simulated the Q-learning model with a random sequence of blocks. This sequence of blocks was independent of the actual sequence of blocks used in that session both in the reward probabilities and in the timing of transition between blocks. To allow for regression of the entire spike sequence (mean and standard deviation of number of trials was 518 and 122, respectively) on the estimated Qs, longer sessions than those of Fig. 1 were simulated. These simulated sessions consisted of 3 random repetitions of the 4 blocks of used in Fig. 1 and were then truncated to fit the length of the spike sequence. As before, we used the results of this simulation to extract estimates of the two action-values and we regressed the sequence of spike counts on these randomly assigned estimated Q-values. The t-values of 642 regressions (214 neurons in three sessions) are presented in Fig. 5a. The standard analysis identified 43% of the neurons as action-value neurons, despite the fact that these action-values were unrelated to the experimental session in which these neurons were recorded (Fig. 5b).

Alternative methods for identifying action-value neurons

So far, our population analysis was based on fitting Eqs. (1) and (2) to the sequence of actions and rewards and using the resultant Q-values as estimates of the action-values (Thin lines in Fig. 1c). A standard alternative approach to estimating action-values, which is model-free, is to use the average reward associated with the block as a measure of the action-value and regress the spike count at the end of the block on it. This is similar to the analysis in the individual examples of Figs. 2a, 3a and 4a (in which two rank sum tests, and not regression, were used)[4]. However, because this analysis is also based on the assumption of independence of trials, applying it to the random-walk neurons, as well as to the experimentally measured neurons, we identify a comparable number of action-value neurons to the one reported in the striatum (Fig. S2).

A previous study has noted that processes such as slow drift in firing rate may violate the independence assumption of the statistical tests[5]. Therefore, they proposed a permutation test, in which the spike count, permuted within each block is also regressed on the estimated Q-values for a large number of different permutations. A neuron is considered an action-value modulated neuron if the t-value of the regression coefficient of the original spike count is large (in absolute value) relative to the distribution of t-values of the permuted spike counts. Using a slightly modified reward schedule[5], it was identified that 10%-13% of striatal neurons were significantly modulated by at least one of the action-values.

However, this method may erroneously identify action-value neurons because the permutation reduces the correlations between the firing rates in consecutive trials. As a result, if there are temporal correlations in the original sequence of spike counts, the regression coefficients for the permuted spike counts are expected to be smaller than that of the original spike count. Indeed, conducting this analysis on the random-walk, the motor cortex, the auditory cortex and the basal ganglia neurons, we found action-value neurons in a comparable number to that reported for the striatum (Fig. S3).

To account for an increasing or decreasing trend in the firing rate (which would result in temporal dependencies in the firing rates), detrending has been applied[8]. In detrending, trial number is added to the regression model as an additional variable. However, this does not remove many of the temporal correlations and indeed, applying the analysis in[8], a comparable number of action-value neurons was found in the random-walk, the motor cortex, the auditory cortex and the basal ganglia neurons to that found in the striatum (Fig. S4).

It has been noted that the significance analyses depicted above are biased towards classifying neurons as action-value neurons, at the expense of state or policy neurons[22]. The reason is that the

former class requires a single significant regression coefficient whereas the latter require two significant regression coefficients (Figs. 1d, 2b, 3b, 4b, 5a). Therefore, an unbiased alternative has been proposed[22]. However, for the same reasons (neural activity in consecutive trials are correlated), this analysis yields a comparable number of action-value neurons in the random-walk, the motor cortex, the auditory cortex and the basal ganglia neurons to that reported in the striatum (Fig. S5).

Additional potential pitfalls and a possible solution

It is tempting to believe that adding more trials in a block or adding more blocks to the experiment may solve the problem of erroneously identifying action-value neurons. The reason is that the larger the number of trials, the less likely it is that a neuron that is not modulated by an 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. We find that increasing the number of blocks can result in a larger probability of an erroneous identification of a neuron whose firing rate has temporal correlations as an action-value neuron. This is because the estimated variance of regression coefficients is proportional to the inverse of the number of degrees of freedom, which increases with the number of trials. As a result, the significance threshold decreases with the number of trials. Specifically, we simulated the random-walk neurons in an eight-block design, where the four blocks from fig 1a were repeated twice (both times in random permutation). We found that 45%+/-1.6% of these random-walk neurons were classified as action-value neurons. Similarly, when the spike counts of neurons from the motor cortex were regressed on estimated Q-values from the 8-blocks design, 37%+/-5% of these neurons were classified as action-value neurons. The same was done for spike counts of neurons from the auditory cortex (341 estimated Q-values with more than 370 trials did not participate in this analysis) and 24%+/-4.7% of these

neurons were classified as action-value neurons. For the basal ganglia neurons, already longer sessions were used, with 12 blocks in each session, truncated to fit the length of spike sequences.

Adding blocks can be useful, however, if the reward schedules in the different blocks are independent, and the number of assumed degrees of freedom in the statistical analysis depends on the number of blocks and not on the number of trials. For example, the single-neuron statistical analysis in Figs. 2a, 3a and 4a is flawed because the variance in the mean spike count (in the last 20 trials of the block) is estimated assuming that the spike counts in consecutive trials are independent from each other. A correct analysis would have considered this mean as a single data point, in which the variance cannot be estimated. However, this is experimentally difficult because it requires a substantially larger number of blocks and thus trials in an experiment, than is typically used.

Our results so far have demonstrated that the regression coefficients of the spike count on Q-values can appear significant even if the neurons are not modulated by action-values. Here we propose a way of identifying action-value neurons, which is not prone to this caveat, and is also applicable to the neurons whose activity was measured in previous publications[4,5,8]. We propose a permutation test, in which we seek neurons that are more correlated with the action-value that was estimated from the session in which the neuron was recorded than with surrogate action-values that were estimated from other sessions. This is illustrated in Fig. 6. First, we computed the t-values of the regression coefficients of the spike counts of the two simulated action-value neurons in Fig. 1b on each of the estimated Q-values from all sessions. Because the number of trials differed between the sessions, we only considered the 504 sessions longer or equal to 170 trials and for these sessions, we used only the first 170 trials. The two distributions of t-values, one for each simulated neuron, are depicted in Fig. 6a. Note that the 5% significance boundaries, which are

exceeded by exactly 5% of t-values in each distribution, are substantially larger (in absolute value) than 2 (1.96 is 97.5th percentile in a t-distribution). There are two reasons for these wide distributions of t-values. First, there are trial-to-trial correlations both in the estimated Q-values and in the spike counts of the simulated action-value neurons. As a result, the effective number of degrees of freedom is substantially smaller than the number of trials, leading to larger t-values than expected when trials are independent. Second, some of the surrogate sessions corresponded to an identical or opposite sequence of reward probabilities, resulting in surrogate estimated Q-values that are highly correlated (or anti-correlated) with the Q-value that is estimated from the neuron's session. We posit that a regression coefficient is significant if the t-value of the regression on the action-value that is estimated from the neuron's session exceeds the significance boundaries derived from the permutations. Indeed, when considering the Top (red) simulated action 1-value neuron, we find that its spike count is significantly correlated with the estimated Q_1 from its session (red arrow) but not with that of estimated Q_2 (blue arrow). 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.6b, black), demonstrating that this analysis can identify action-value neurons. When considering the random-walk neurons (Fig. 2) or two of the experimentally measured neurons (Figs. 3 and 4), this method defines only approximately 10% of the neurons as action-value neurons, as predicted by chance (Fig. 6b).

Applying this method to the basal ganglia neurons with Q-values estimated from the free-choice sessions in[21] (unlike the simulated Q-values used in Fig. 5), we find that the original analysis of

regression on estimated Q-values results in 16% $(103/(214 \times 3))$ and 16% $(100/(214 \times 3))$ of the neurons × phases represent the left and right action-values, respectively (with $p<0.01$). By contrast, the permutation analysis yields that only 1.9% $(12/(214 \times 3))$ and 1.7% $(11/(214 \times 3))$ of the neurons × phases have a significantly higher regression coefficient with their corresponding left or right action-values, respectively, than with surrogate action-values (Fig. 6c).

Two points are noteworthy regarding this alternative analysis. First, Fig. 6a demonstrates that the distribution of the t-values of the regression of the spike count of a neuron on all action-values depends on the neuron. 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 per session is different between sessions. Second, this analysis is still biased towards classifying neurons as action-value neurons at the expense of state or policy neurons, as noted above[22]. Therefore, it may erroneously identify neurons whose activity is correlated with other decision variables, such as state or policy, as action-value neurons (Fig. S6).

Discussion

In this paper, we showed that temporal dependencies in the firing rates of neurons may result in their erroneous identification as representing action-values when utilizing standard analysis methods. Specifically, we considered a particular example of a violation of the independence of trials by simulating neurons whose firing rate follows a bounded random-walk process and erroneously identified action-value representations in these neurons. Moreover, we erroneously identified "action-value" neurons recorded in different cortical regions in unrelated experiments.

Furthermore, we erroneously identified representations of unrelated "action-values" in basal ganglia neurons. Finally, we proposed an alternative method of analysis that is not subject to this limitation, which can be utilized to reanalyze previous experiments. Applying this novel method to basal ganglia recordings, in which action-value neurons were previously identified, we failed to detect action-value representations.

It is important to note that we do not take these results to imply that erroneous detection of action-value representation may occur in every brain region. On the contrary, neurons in different brain areas and within the same brain area differ according to their degree of non-stationarity and the time-constants of the slow modulations that cause this non-stationarity, features which affect the probability of erroneous detection of action-value representation. Further, neurons that represent features correlated with action-value, such as choice, are more likely to be classified as action-value neurons (see Fig. S6). On a similar note, these results do not imply that action-value representation should be found regardless of the experimental epoch analyzed, as degree of non-stationarity may also change in time. For example, if a neuron represents an unrelated slowly changing variable only in a certain time epoch in the trial it may be erroneously classified as an action-value neuron, but only when considering this specific time epoch.

These results have direct consequences regarding action-value representations in the striatum. To the best of our knowledge, all electrophysiological studies that provide direct evidence that neurons in the striatum are specifically modulated by action-values are susceptible to the caveat discussed in this paper[4,5,7,8,10,22]. In view of these findings, the well-accepted hypothesis that neurons in the striatum encode action-values should be reconsidered.

Beyond this potential pitfall in the statistical analyses of neuronal recordings, it has been noted that identifying a neuronal correlate of value is difficult, because it is hard to untangle value from

other variables, such as salience, the outcome's sensory properties or information about properties of the task[23]. Moreover, value is an internal variable that is crucially dependent on the specific model used to define it. Finally, in the common experimental settings where a subject chooses between different options with different values it is typically impossible to differentiate direct-policy signals from value signals[24]. This is because the policy variable will be similar to some function of the difference between the action-values and therefore, depending on the algorithm, may correlate strongly with an action-value, the chosen value or prediction error.

The potential statistical pitfalls associated with non-stationarity of neural activity are relevant to any attempt to identify the neural correlate, be it spike count of a neuron or BOLD signal at a voxel, of a slowly changing variable. Our focus here was action-value representation by single neurons, but other variables associated with gradual learning are also likely to be slowly varying and hence identifying them using any measure of neural activity will pose a similar challenge. Another place where such a problem may arise is in the estimation of noise correlations. For example, a population of neurons whose firing rates follow independent random-walk processes (or any other temporally correlated process), may appear to be endowed with significant noise correlations. Similar issues may arise when studying correlations in BOLD signal between voxels when studying resting state functional connectivity. Any statistical test performed in these cases should consider the possibility of irrelevant non-stationarity of the activity.

The fact that the basal ganglia in general and the striatum in particular play an important role in operant learning, planning and decision-making is indisputable[3,25–29]. However, our results show that special caution should be applied when relating the activities of neurons there with internal variables derived from specific reinforcement learning algorithms. Especially, the generally held hypothesis that neurons in the striatum represent action-values should be reconsidered.

Materials and Methods

The action-value neurons model

To model neurons whose firing rate is modulated by an action-value, we considered neurons whose firing rate changes according to:

$$f(t) = B + K \cdot r \cdot (Q_i(t) - 0.5) \quad (4)$$

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.

Estimation of Q-values from model choices and rewards

To imitate experimental procedures, we regressed the spike count on estimates of the Q-values, rather than the Q-values that underlied behavior (to which the experimentalist has no direct access). For that goal, for each session, we assumed that $Q_i(1) = 0.5$ and found the set of parameters $\hat{\alpha}$ and $\hat{\beta}$ that yielded the estimated Q-values that best fit the sequences of actions in each experiment by maximizing the likelihood of the sequence. Q-values were estimated from Eq. (1), using these estimated parameters and the sequence of actions and rewards. Overall, the estimated values of the parameters α and β 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 α=0.1 and β=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. 34% (42/125) of the auditory cortex neurons were excluded on average and 23% (29/125) were excluded in all 40 repetitions. 20% (126/(214×3)) of basal ganglia neurons were excluded on average and 11% were excluded in all 40 repetitions (74/(214×3)). None of the simulated action-value neurons (0/20,000) or the motor cortex neurons (0/89) was excluded.

Statistical analyses

The computation of the t-values of the regression of the spike counts on the estimated Q-values was done using *regstats* in MATLAB. The following regression model was used:

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

Where $\mathrm{s(t)}$ 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 β_{0-2} are the regression parameters.

To find neurons whose spike count in the last 20 trials is modulated by reward probability (Figs. 1b, 2a, 3a, 4a), we executed the Wilcoxon rank sum test, using *ranksum* in MATLAB. All tests were two-tailed.

The motor cortex recordings

The data 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 BMI and sensorimotor combined task. Spikes and LFP 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 move 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[30], $r_{SC} = 0.047 \pm 0.17$ (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

The auditory cortex recordings are described in detail in[20]. In short, membrane potential was recorded intracellularly from 39 neurons in the auditory cortex of anesthetized rats. 125 experimental session 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

The basal ganglia recordings are described in detail in[21]. 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 sessions there were no neural recordings).

Data 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. Other data are available upon request.

Code Availability

Custom MATLAB scripts used to create simulated neurons and to analyze data are also available upon request.

Author Contributions

L.E.D. and Y.L. designed the analysis and wrote the paper

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 and Roey Schurr for careful reading of the

manuscript and helpful comments and Inbal Goshen and Hanan Shteingart for discussions. This work was supported by the Israel Science Foundation (Grant No. 757/16), CRC1080 and the Gatsby Charitable Foundation.

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).

2. Louie, K. & Glimcher, P. W. Efficient coding and the neural representation of value. *Ann. N. Y. Acad. Sci.* **1251,** 13–32 (2012).

3. Schultz, W. Neuronal Reward and Decision Signals: from Theories to Data. *Physiol. Rev.* **95,** 853–951 (2015).

4. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of Action-Specific Reward Values in the Striatum. *Science* **310,** 1337–1340 (2005).

5. Kim, H., Sul, J. H., Huh, N., Lee, D. & Jung, M. W. Role of Striatum in Updating Values of Chosen Actions. *J. Neurosci.* **29,** 14701–14712 (2009).

6. Shteingart, H. & Loewenstein, Y. Reinforcement learning and human behavior. *Curr. Opin. Neurobiol.* **25,** 93–98 (2014).

7. Lau, B. & Glimcher, P. W. Value Representations in the Primate Striatum during Matching Behavior. *Neuron* **58,** 451–463 (2008).

8. Ito, M. & Doya, K. Distinct Neural Representation in the Dorsolateral, Dorsomedial, and Ventral Parts of the Striatum during Fixed- and Free-Choice Tasks. *J. Neurosci.* **35,** 3499–3514 (2015).

9. Funamizu, A., Ito, M., Doya, K., Kanzaki, R. & Takahashi, H. Condition interference in rats performing a choice task with switched variable- and fixed-reward conditions. *Front. Neurosci.* **9,** 1–14 (2015).

10. Ito, M. & Doya, K. Parallel Representation of Value-Based and Finite State-Based

Strategies in the Ventral and Dorsal Striatum. *PLoS Comput. Biol.* **11,** 1–25 (2015).

11. Seung, H. S. Learning in Spiking Neural Networks by Reinforcement of Stochastic Synaptic Transmission. *Neuron* **40,** 1063–1073 (2003).

12. Fiete, I. R., Fee, M. S. & Seung, H. S. Model of Birdsong Learning Based on Gradient Estimation by Dynamic Perturbation of Neural Conductances. *J. Neurophysiol.* **98,** 2038–2057 (2007).

13. Urbanczik, R. & Senn, W. Reinforcement learning in populations of spiking neurons. *Nat. Neurosci.* **12,** 250–252 (2009).

14. Darshan, R., Leblois, A. & Hansel, D. Interference and Shaping in Sensorimotor Adaptations with Rewards. *PLoS Comput. Biol.* **10,** 1–20 (2014).

15. Neiman, T. & Loewenstein, Y. Covariance-Based Synaptic Plasticity in an Attractor Network Model Accounts for Fast Adaptation in Free Operant Learning. *J. Neurosci.* **33,** 1521–1534 (2013).

16. Fremaux, N., Sprekeler, H. & Gerstner, W. Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity. *J. Neurosci.* **30,** 13326–13337 (2010).

17. Loewenstein, Y. & Seung, H. S. Operant matching is a generic outcome of synaptic plasticity based on the covariance between reward and neural activity. *PNAS* **103,** 15224–15229 (2006).

18. Loewenstein, Y. Robustness of Learning That Is Based on Covariance- Driven Synaptic Plasticity. *PLoS Comput. Biol.* **4,** 1–10 (2008).

19. Loewenstein, Y. Synaptic theory of Replicator-like melioration. *Front. Comput. Neurosci.*

4, 1–12 (2010).

20. Hershenhoren, I., Taaseh, N., Antunes, F. M. & Nelken, I. Intracellular Correlates of Stimulus-Specific Adaptation. *J. Neurosci.* **34,** 3303–3319 (2014).

21. Ito, M. & Doya, K. Validation of Decision-Making Models and Analysis of Decision Variables in the rat basal ganglia. *J. Neurosci.* **29,** 9861–9874 (2009).

22. Wang, A. Y., Miura, K. & Uchida, N. The dorsomedial striatum encodes net expected return, critical for energizing performance vigor. *Nat. Neurosci.* **16,** 639–47 (2013).

23. O'Doherty, J. The problem with value. *Neurosci. Biobehav. Rev.* **43,** 259–268 (2014).

24. Li, J. & Daw, N. D. Signals in Human Striatum Are Appropriate for Policy Update Rather than Value Prediction. *J. Neurosci.* **31,** 5504–5511 (2011).

25. McDonald, R. . J. & White, N. . M. A Triple Dissociation of Memory Systems: Hippocampus, Amygdala, and Dorsal Striatum. *Behav. Neurosci.* **107,** 3–22 (1993).

26. O'Doherty, J. *et al.* Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science* **304,** 452–454 (2004).

27. Thorn, C. A., Atallah, H., Howe, M. & Graybiel, A. M. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops During Learning. *Neuron* **66,** 781–795 (2010).

28. Yarom, O. & Cohen, D. Putative cholinergic interneurons in the ventral and dorsal regions of the striatum have distinct roles in a two choice alternative association task. *Front. Syst. Neurosci.* **5,** 1–9 (2011).

29. Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of

distinct subpopulations of striatal neurons mimics changes in action value. *Nat. Neurosci.* **15,** 1281–9 (2012).

30. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14,** 811–819 (2011).

Figures

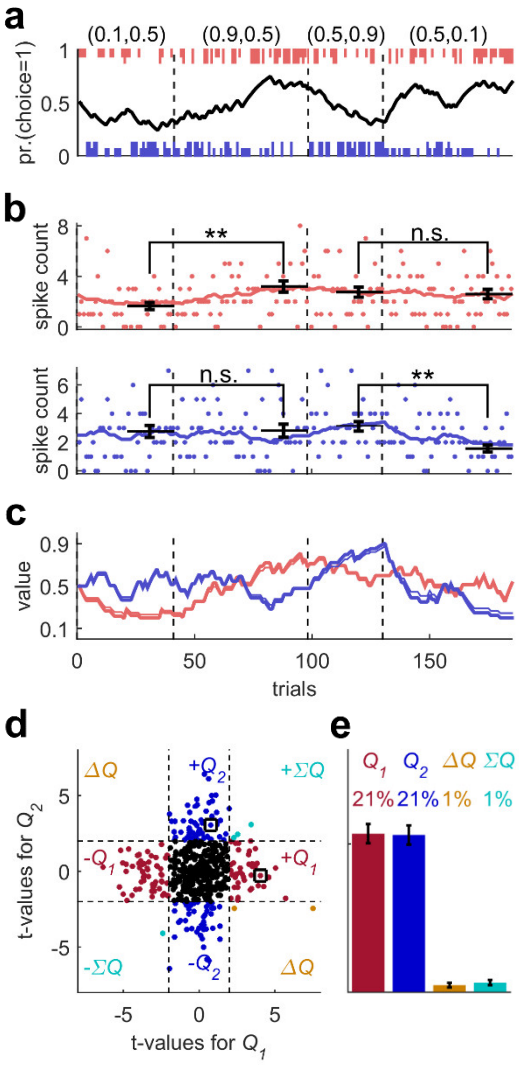

Figure 1 Model of action-value neurons **(a)** Behavior of model in example session, composed of four blocks (separated by a dashed vertical line). The probabilities of reward for choosing actions 1 and 2 are denoted by the pair of numbers above the block. Black line denotes the probability of choosing 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 estimations of Q_1 and Q_2, respectively, based on the choices and rewards in A. 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. Color of dots denote significance: dark red and blue denote significant regression coefficient only on one estimated action-value, action 1 and action 2, respectively; light blue – significant regression coefficients on both estimated action-values with similar signs (ΣQ), orange - significant regression coefficients on both estimated action-values with opposite signs (ΔQ). 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.

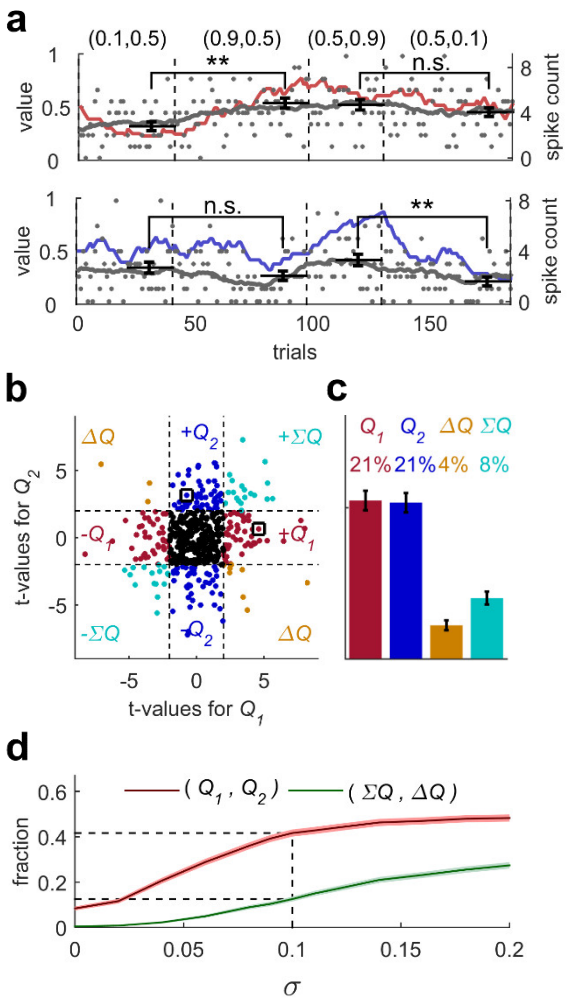

Figure 2 Erroneous detection of action-value representation in random-walk neurons **(a)** Two example random-walk neurons presented as if the sequence of spike counts of these neurons corresponds to the sequence of trials in figure 1. Gray lines and gray dots denote the neurons' firing rates and the spike counts, 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). The red and blue lines denote the estimated action-values 1 and 2, respectively (same as in Fig. 1c). These are presented here because the t-value of the regression of the spike counts on the corresponding estimated action-values was larger than 2. **(b)** and **(c)** Population analysis, same as in Figs.1d and 1e for the random-walk neurons. The two random-walk neurons in **(a)** are denoted by squares in **(b)**. **(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 Fig. 2a-c. Initial firing rate for all neurons in the simulations is $f(1) = 2.5$Hz.

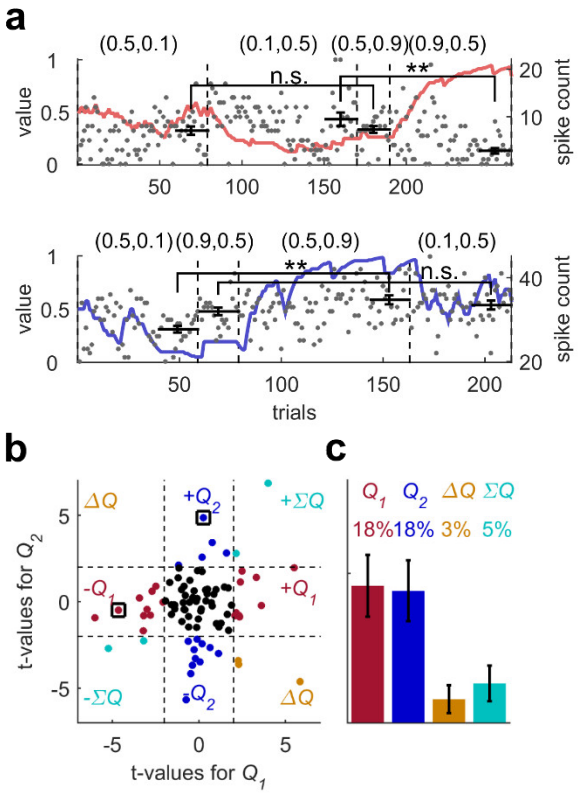

Figure 3 Erroneous detection of action-value representation in motor cortex **(a)** Two example motor cortex neurons recorded in a BMI task, presented as if the sequence of spike counts of these neurons corresponds to the sequence of trials in two sessions (one for each neuron) of operant learning used for the population analysis in Fig. 1e. Gray dots denote the spike-counts. Black horizontal lines denote 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). Each session was associated with two estimated action-values and for each neuron, we computed the t-values of the regression of the spike counts on the two corresponding estimated action-values. The red and blue lines denote those action-values whose t-value exceeded 2 (in absolute value). **(b)** and **(c)** Population analysis. **(b)** The t-values of 89 neurons regressed on the estimated action-values of randomly selected 89 sessions (same as Fig. 1d). The neurons in **(a)** are denoted by squares. **(c)** 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. Error bars denote the standard error of the mean.

.

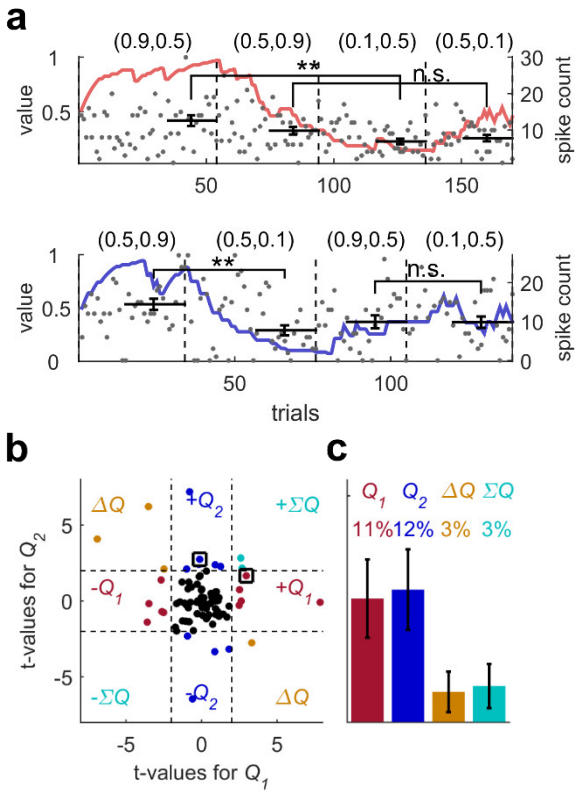

Figure 4 Erroneous detection of action-value representation in auditory cortex **(a)** Same as in Fig. 3a for two auditory cortex neurons in an anesthetized rat responding to the presentation of pure tones. **(b)** and **(c)** Population analysis. **(b)**. The t-values of 82 neurons regressed on the estimated action-values of randomly selected 82 sessions (same as Fig. 3b). The neurons in **(a)** are denoted by squares. **(c)** Fraction of neurons classified in each category, estimated by regressing 125 auditory cortex neurons on 80 different estimated action-values from 40 randomly selected sessions (in each session, 34% of neurons were excluded on average, see Materials and Methods). Error bars denote the standard error of the mean.

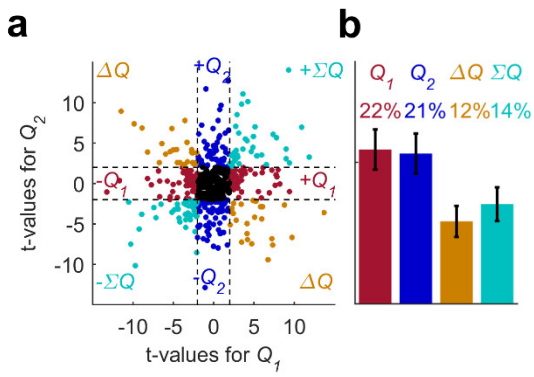

Figure 5 Erroneous detection of irrelevant action-value representation in basal ganglia **(a)** and **(b)** Population analysis. **(a)** The t-values of 214 neurons in three different phases regressed on the estimated action-values from randomly selected 642 simulated sessions (same as Fig. 4b). **(b)** 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 (in each session, 20% of neurons were excluded on average, see Materials and Methods). Error bars denote the standard error of the mean.

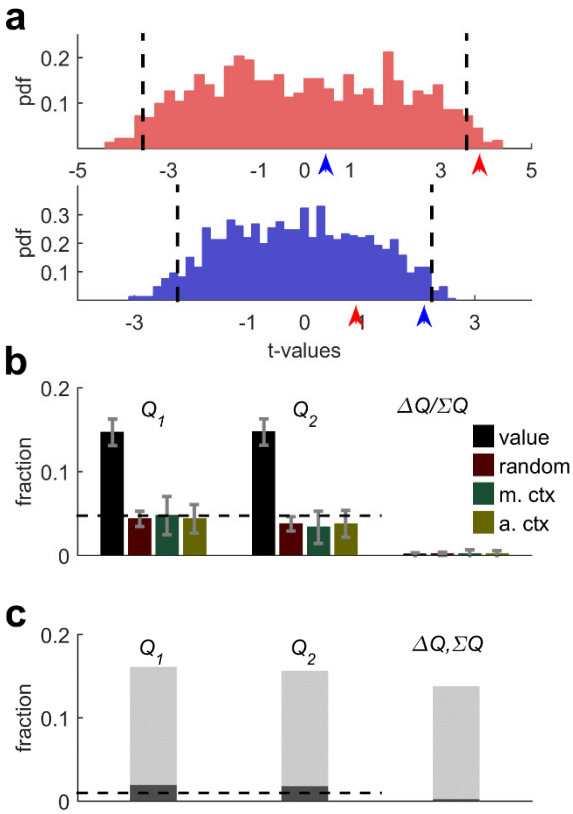

Figure 6 Permutation analysis **(a)** Red and blue figures correspond to red and blue - labeled neurons in fig. 1b, respectively. For each neuron, we computed the t-values of the regressions of its spike-count on both estimated Q-values from all sessions in Fig. 1e (excluding sessions shorter than 170 trials) and used these t-values to compute probability distribution functions of the t-values. Dashed black lines denote the 5% significance boundary. Red and blue arrows denote the t-values from regressions on the estimated Q_1 and Q_2, respectively, from the session in which the neuron was simulated (depicted in Fig. 1a). **(b)** Fraction of neurons classified in each category using the permutation analysis for the action-value neurons (black, Fig. 1), random-walk neurons (maroon, Fig. 2), motor cortex neurons (green, Fig. 3) and auditory cortex neurons (dark yellow, Fig. 4). Dashed line denotes chance level for action-value 1 or 2 classification. Error bars denote the standard error of the mean. **(c)** Light gray, fraction of basal ganglia neurons classified in each category when regressing the spike count of basal ganglia neurons on the estimated Q-values associated with their experimental session. Dark gray, fraction of basal ganglia neurons classified in each category when applying the permutation analysis. Dashed line denotes significance level of $p<0.01$.