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 action. 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;

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 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[1]. In recent years, there is a lot of interest in the neural mechanisms underlying this computation[2,3]. In particular, based on electrophysiological experiments[4–15], 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[4,6]. 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 sigmoidal function, typically softmax, of the difference of the action-values (see also[16]):

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

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 and similar results were considered as evidence that neurons in the striatum represent action-values[4–10,12].

In this paper we point out that this literature has widely ignored a known caveat in regression analysis - it can result in erroneous identification of neurons as representing action-value if the

66 firing rates are temporally correlated. After a systematic literature search we conclude that this
67 caveat has not yet been fully addressed. We maintain that the conclusion that there is representation
68 of action-value in the striatum must await new evidence that is not prone to this caveat.

69 Three clarifications are required. First, although this paper discusses a methodological problem
70 that may also be of relevance in other fields of neuroscience, we focus on a single scientific claim,
71 namely that a representation of action-values in the striatum is an established fact. Second, our
72 criticism is restricted to the representation of action-values, and we do not make any claims
73 regarding the possible representations of other decision variables, such as policy, chosen-value or
74 reward-prediction-error. Third, we focus on the striatum and do not make claims about the possible
75 representations of action-values elsewhere in the brain.

76 The paper is organized in the following way. We commence by describing a standard method for
77 identifying action-value neurons. Then, we demonstrate that this method erroneously identifies
78 action-value neurons when they do not exist in a mathematical model, as well as in unrelated
79 neuronal recordings from the motor cortex of a monkey, the auditory cortex of anaesthetized rats
80 and the basal ganglia of behaving rats. Finally, we conduct a systematic literature search and show
81 that all alternative approaches for identifying action-value neurons that were previously used can
82 also lead to the erroneous identification of action-value neurons. We conclude by proposing a
83 different method for identifying action-value neurons, that is not subject to this caveat and applying
84 it to basal ganglia recordings, in which action-value neurons were previously identified. Using this
85 new method, we fail to detect any action-value representations.

86 **Results**

87 **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[4,10].

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[6]) 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[4–6,10–12]. 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. 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. Such neurons are typically classified as state or policy (preference) neurons, if the two regression coefficients have the same or different signs, respectively[10]. 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 Q-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.

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 the 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 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 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 as a function of the diffusion parameter σ (Fig. 2d). When σ=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%, 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 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 temporally-proximate trials, violating the independence-of-trials assumption of the test. Because of these dependencies, the expected variability in 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 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-of-trials 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[17–22]. 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[21,23–25] 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. This was a brain-machine-interface (BMI) experiment composed of 600 identical trials (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[26] (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[7]. 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 reward schedule in each block was pseudo randomly chosen from the ones presented in Fig. 1a. Blocks changed when the subject chose the higher-valued action at least 80% of the trials within 20 trials. As in[7], 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 Q-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 completely unrelated to the experimental session in which these neurons were recorded (Fig. 5b).

Alternative approaches 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).

Our analyses in Figs. 2-5 clearly demonstrate that this approach can lead to the erroneous identification of action-values neurons. However, other studies have concluded that action-value is represented in the striatum when utilizing alternative approaches. In order to challenge the finding of action-value neurons in the striatum, we conducted a literature search to find all the alternative approaches used to identify action-value representation in the striatum (see Materials and Methods). 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, 3 papers have used the term action-value to refer to the value of the *chosen* action (also known as chosen-value)[27–29] and therefore we will not discuss them any further.

A second group of 11 papers did not distinguish between action-value and policy representations[5,6,9,12–14], or reported policy representation[15,30–33]. While action-value representation is often implied from policy representation, it is well-known that policy representation can emerge in the absence of action-value representation. For example, in computer science, direct policy-gradient methods that do not entail values are routinely used[34]. In neuroscience, several studies have proposed neuronal mechanisms that approximate direct policy reinforcement learning and decision making by means of reward-modulated synaptic plasticity (e.g., [17–25]). All these models will result in policy representation in the absence of action-value representation. For this reason, these findings do not necessarily imply action-value representation in the striatum.

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[35,36]. However, a change in policy does not entail an action-value representation because, as noted above, a policy can be learned (or preference emerge) and be modulated by reinforcers without any action-value

computation. Therefore, these papers are not a 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[4,7,8,10,11,37]. Five used electrophysiological recordings of single units in the striatum and one was an fMRI study. All used block-design experiments where action-values are temporally correlated. In addition to the regression of the spike count on estimated Q-values described in Figs. 1-5 and S1, some of them considered alternative approaches. However, as described below, these alternatives are subject to the same caveat.

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[4,7]. This is similar to the analysis in the individual examples of Figs. 1b, 2a, 3a, 4a and S1b (in which two rank sum tests, and not regression, were used). 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).

In principle, temporal dependencies in the firing rates could result from trends. Indeed, detrending has been applied to the spike count[10]. In detrending, trial number is added to the regression model as an additional variable. However, this does not remove many of the temporal correlations. Indeed, we find a comparable number of erroneously-identified action-value neurons to that found in the striatum[10] when applying this analysis to the random-walk model, the motor cortex, the auditory cortex and the basal ganglia neurons (Fig. S3).

It has also 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[8]. 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, S1b, S2 and S3). Therefore, an unbiased alternative has been proposed[8]. However, for the same reasons (neural activities in consecutive trials are correlated), this analysis yields a comparable number of erroneously-identified 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. S4).

Taken together, we conclude 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.

Attempts to account for the non-stationarity

The mean length of the sessions used in the analysis in Figures 1-4 was 174 trials (standard deviation 43 trials). It is tempting to believe that adding more trials in a block or adding more blocks to the experiment may solve the problem of erroneous identification of action-value neurons. The idea 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. Specifically, we increased the number of trials by simulating the random-walk neurons in an eight-block design (as opposed to 4 blocks in Fig. 1), where the four blocks from fig 1a were repeated twice (both times in random permutation). The resulting mean length of the sessions was 347 trials (standard deviation 65 trials). We found that 45%±1.6% of the 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 (324 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. We did not perform this analysis on the basal ganglia neurons because we were limited by the length of these recordings.

In fact, our results suggest that increasing the number of blocks can result in a larger fraction of erroneously-identified action-value neurons. 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.

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 of 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.

Trial design experiments are not prone to the caveat we discuss here because by construction, the different trials are independent from each other, so the predictors in consecutive trials are not correlated. However, learning the values of actions requires that reward probabilities (or

359 magnitudes) in consecutive trials will be strongly correlated, which is not possible in trial design.
360 Several studies used tasks, in which cues mark the reward-probability[14,15,33]. This way it is possible
361 to use a trial design, in which the expected rewards associated with an action in consecutive trials
362 are independent. However, these studies did not distinguish between values and policy
363 representations. It should be noted that in this design, the learning is the association of cue and
364 reward (cue-values), and not the association of action and reward (action-values)[38].

365 Two studies noted that processes such as slow drift in firing rate may violate the independence-of-
366 trials assumption of the statistical tests and suggested unique methods to deal with this issue in a
367 block-design[6,9]. Although in these studies action-value representation was not differentiated from
368 policy representation, we repeated the methods described in them to see if they are subject to the
369 same caveat described above. As shown below, we report erroneous detection of action-value
370 neurons even when these methods are applied.

371 In the first study[6], a permutation test was proposed, in which the spike count, permuted within
372 each block is also regressed on the estimated Q-values for a large number of different permutations.
373 A neuron is considered as an action-value modulated neuron if the t-value of the regression
374 coefficient of the original spike count is large (in absolute value) relative to the distribution of t-
375 values of the permuted spike counts. Using a slightly modified reward schedule[6], it was identified
376 that 10%-13% of striatal neurons were significantly modulated by at least one of the action-values.

377 However, this method may erroneously identify action-value neurons because the permutation
378 reduces the correlations between the firing rates in consecutive trials. As a result, if there are
379 temporal correlations in the original sequence of spike counts, the regression coefficients for the
380 permuted spike counts are expected to be smaller than that of the original spike count. Indeed,
381 conducting this analysis on the random-walk, the motor cortex, the auditory cortex and the basal

382 ganglia neurons, we found action-value neurons in a comparable number to that reported for the
383 striatum (Fig. S5)

384 In the second study[9], the spike-counts in the last 3 trials were also used as predictors in the
385 regression model. However, this method does not address longer term dependencies and
386 conducting this analysis using the same reward schedule as in[9] on the random-walk, the motor
387 cortex, the auditory cortex and the basal ganglia neurons, we found action-value neurons in a
388 comparable number to that reported for the striatum (Fig. S6).

389 **A possible solution to statistical significance in non-stationary time-series**

390 The concerns about the statistical tests described above result from the possibility that non-
391 stationarity of the spike count in striatal neurons is not the result of action-value learning. In
392 principle, if the statistical structure of this non-stationarity is known, it may be possible to construct
393 a statistical test, such as generalized least squares (GLS)[39] that decorrelates the trials. Alternatively,
394 it may be possible to compare the number of identified action-value neurons to the expected
395 number of erroneously detected "action-value" neurons. However, both approaches require an
396 accurate model of the learning-independent non-stationarity, which is absent, and cannot be
397 extracted from global measures such as the autocorrelation. This is because the non-stationarity
398 could be due to many different processes, including oscillations, trends, random-walks and their
399 combinations.

400 Therefore, we propose a non-parametric permutation test that does not make specific assumptions
401 about the learning-independent non-stationarity. In contrast to the previously-described methods,
402 this test allows us to estimate the probability of an erroneous detection of an action-value neuron

403 under the null hypothesis. Importantly, this method can also be used to reanalyze the activity of
404 previously-recorded striatal neurons[4–10].

405 We propose a permutation test, in which we seek neurons that are more correlated with the action-
406 value that was estimated from the session in which the neuron was recorded than with surrogate
407 action-values that were estimated from other sessions. This is illustrated in Fig. 6. First, we
408 computed the t-values of the regression coefficients of the spike counts of the two simulated action-
409 value neurons in Fig. 1b on each of the estimated Q-values from all relevant sessions (see below).
410 The two distributions of t-values, one for each simulated neuron, are depicted in Fig. 6a. Note that
411 the 5% significance boundaries, which are exceeded by exactly 5% of t-values in each distribution,
412 are substantially larger (in absolute value) than 2 (1.96 is 97.5th percentile in a t-distribution). There
413 are two reasons for these wide distributions of t-values. First, there are trial-to-trial correlations
414 both in the estimated Q-values and in the spike counts of the simulated action-value neurons. As a
415 result, the effective number of degrees of freedom is substantially smaller than the number of trials,
416 leading to larger t-values than expected when trials are independent. Second, some of the surrogate
417 sessions corresponded to an identical or opposite sequence of reward probabilities, resulting in
418 surrogate estimated Q-values that are highly correlated (or anti-correlated) with the Q-value that
419 is estimated from the neuron's session. We posit that a regression coefficient is significant if the t-
420 value of the regression on the Q-value that is estimated from the neuron's session exceeds the
421 significance boundaries derived from the permutations. Indeed, when considering the Top (red)
422 simulated action 1-value neuron, we find that its spike count is significantly correlated with the
423 estimated Q_1 from its session (red arrow) but not with that of estimated Q_2 (blue arrow). Because
424 the significance boundary exceeds 2, this approach is less sensitive than the original one (Fig. 1)
425 and indeed, the regression coefficients of the Bottom simulated neuron (blue) do not exceed the

426 significance level (red and blue arrows) and thus this analysis fails to identify it as an action-value
427 neuron.

428 Considering the population of simulated action-value neurons of Fig. 1, this analysis identified
429 29% of the action-value neurons of Fig. 1 as such (Fig.6b, black), demonstrating that this analysis
430 can identify action-value neurons. When considering the random-walk neurons (Fig. 2) or two of
431 the experimentally measured neurons (Figs. 3 and 4), this method defines only approximately 10%
432 of the neurons as action-value neurons, as predicted by chance (Fig. 6b).

433 One technical point of caution is that the number of trials can affect the distribution of t-values.
434 Therefore, we only considered in our analysis the first 170 trials of the 504 sessions longer or equal
435 to 170 trials.

436 We used this new method to consider action-value representation in the basal ganglia. To that goal,
437 we considered the recordings reported in [7]. That paper utilized the model-free method depicted in
438 Fig. S2 to identify action-value neurons. They reported that 13% and 10% of the neurons × phases
439 represent the left and right action-values, respectively (with $p<0.01$), suggesting that
440 approximately 23% of the striatal neurons represent action-values at different phases of the
441 experiment. As a first step in our analysis, we applied the standard model-based approach
442 presented in Figs. 1d, 2b, 3b, 4b, 5a: we used the behavior of the animals to estimate the Q-values
443 and regressed the spike counts in the three phases of the experiment on the estimated Q-values.
444 This analysis yielded that approximately 32% of the neurons represent action values (16%
445 $(103/(214 \times 3))$ and 16% $(100/(214 \times 3))$ of the neurons × phases represent the left and right
446 action-values, respectively with $p<0.01$), a number that is slightly higher than the result of the
447 model-free approach (Fig. 6c). Next, we applied the permutation analysis. Remarkably, this

448 anlalysis yielded that only 3.6% of the neurons (1.9% ($12/(214 \times 3)$) and 1.7% ($11/(214 \times 3)$)
449 of the neurons × phases) have a significantly higher regression coefficient with their corresponding
450 left or right action-values, respectively, than with surrogate action-values (Fig. 6c). These results
451 further challenge the hypothesis of action-value representation in the striatum.

452 It is worth pointing out that the fraction of action-value neurons reported in[7] is low relative to other
453 publications[4,10], a difference that has been attributed to the location of the recording in the striatum
454 (ventral as opposed to dorsal). It would be interesting to apply this method to other striatal
455 recordings[4,8,10].

456 Two points are noteworthy regarding this alternative analysis. First, Fig. 6a demonstrates that the
457 distribution of the t-values of the regression of the spike count of a neuron on all action-values
458 depends on the neuron. Similarly, the distribution of the t-values of the regression of the spike
459 counts of all neurons on an action-value depends on the action-value (not shown). Therefore, the
460 analysis could be biased in favor (or against) finding action-value neurons if the number of neurons
461 per session is different between sessions. Second, this analysis is still biased towards classifying
462 neurons as action-value neurons at the expense of state or policy neurons, as noted above[8].
463 Therefore, it may erroneously identify neurons whose activity is correlated with other decision
464 variables, such as state or policy, as action-value neurons (Fig. S7). To prevent this, it is useful to
465 apply the correction suggested in[8].

466 **Discussion**

467 In this paper, we performed a systematic literature search to discern the methods that have
468 previously been used to infer the representation of action-values in the striatum. We show that
469 none of the methods that have been proposed to distinguish between action-value representation

and other decision variables are able to overcome a serious statistical caveat: temporal dependencies in the firing rates of neurons may result in their erroneous identification as representing action-values. Specifically, we considered a particular example of a violation of the independence-of-trials assumption by simulating neurons whose firing rates follow a bounded random-walk process. We erroneously identified apparent action-value representations in these simulated neurons. Moreover, these methods also erroneously identified neurons recorded in unrelated experiments in different cortical regions, as well as in the basal ganglia, as representing action-values. We propose an alternative method of analysis that is not subject to this limitation, which can be utilized to reanalyze data from previous experiments. When applying this novel method to basal ganglia recordings in which apparent 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 and in any epoch of the trial. On the contrary, neurons in different brain areas, and even within the same brain area, differ according to their degree of non-stationarity and the time-constants of the modulations that cause this non-stationarity; features that both affect the probability of erroneous detection of action-value representation. Indeed, the fraction of erroneously identified action-value neurons differed between the auditory and motor cortices (compare Figs. 3 and 4). Considering the ventral striatum, our analysis indicates that the identification of action-value representations there may have been erroneous, resulting from non-stationary firing rates (Figs. 5, 6c). We were unable to directly analyze recordings from the dorsal striatum because relevant raw data is not publically available. However, previous studies have shown that the firing rates of dorsal-striatal neurons change slowly over time[40,41]. As a result, spike counts are temporally correlated, and violate the independence-

493 of-trials assumption which underlies all previous attempts to identify action-value neurons
494 there[4,8,10,11].

495 The potential statistical pitfalls associated with non-stationarity of neural activity are relevant to
496 attempts to identify any neural correlate of a slowly changing variable, be it the spike count of a
497 neuron, EEG signal from a sensor, or BOLD signal in a voxel. Our focus here was neural
498 representations of action-value, but other variables associated with gradual learning are also likely
499 to vary slowly and hence identifying them using any measure of neural activity will pose a similar
500 challenge.

501 Another situation in which such a problem may arise is in the estimation of noise correlations. For
502 example, a population of neurons whose firing rates follow independent random-walk processes
503 (or any other temporally correlated process), may appear to be endowed with significant noise
504 correlations – again due to the violation of the independence-of-trials assumption, which leads to
505 an underestimation of the variance of the correlation statistic under the null-hypothesis. Similar
506 issues may arise when studying correlations in BOLD fluctuations between voxels when assessing
507 resting state functional connectivity (see[42,43] for discussions of autocorrelation in fMRI analyses).
508 Any statistical test performed in these cases should consider the possibility of irrelevant non-
509 stationarity.

510 Returning to the question of action-value representations in the striatum, it has been previously
511 noted that identifying a specific neuronal correlate of value is difficult, because it is hard to
512 disentangle value from other variables, such as salience, the outcome's sensory properties or
513 information about the properties of the task[44]. It is also difficult to disentangle action-value
514 representation from choice representation, as shown in Fig. S7.

To 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 statistical caveat demonstrated in this paper[4,7,8,10,11,37], or reported results in a manner indistinguishable from policy, which does not necessarily imply value representation (as shown in Fig. S1)[14,15,33]. Indeed, many studies were susceptible to both of these confounds[5,6,9,12,13]. 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[45,46] and one noted that lesions to the dorsal striatum do not impair action-value learning[47].

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 not in question[3,35,48–52]. However, our results show that special caution should be applied when relating activity in neurons there with specific variables, derived from reinforcement learning algorithms, which vary slowly over time. The prevailing belief that neurons in the striatum represent action-values must await further tests that can account for the potential caveats discussed here.

Materials and Methods

Literature search

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 articles. Overall, we identified 22 papers that reported new evidence in support of action-value representation in striatal neurons [4–15,27–33,35–37] and these were considered in this manuscript.

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).

559 *Exclusion of neurons*

560 Following standard procedures, a sequence of spike-counts, either simulated or experimentally
561 measured was excluded due to low firing rate if the mean spike count in all blocks was smaller
562 than 1. This procedure excluded 0.02% (4/20,000) of the random-walk neurons. 34% (42/125) of
563 the auditory cortex neurons were excluded on average and 23% (29/125) were excluded in all 40
564 repetitions. 20% (126/(214×3)) of basal ganglia neurons were excluded on average and 11% were
565 excluded in all 40 repetitions (74/(214×3)). None of the simulated action-value neurons (0/20,000)
566 or the motor cortex neurons (0/89) was excluded.

567 *Statistical analyses*

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

570 $$s(t) = \beta_0 + \beta_1 Q_1(t) + \beta_2 Q_2(t) + \epsilon(t)$$

571 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,
572 $\epsilon(t)$ is the residual error in trial t and β_{0-2} are the regression parameters.

573 To find neurons whose spike count in the last 20 trials is modulated by reward probability (Figs.
574 1b, 2a, 3a, 4a), we executed the Wilcoxon rank sum test, using *ranksum* in MATLAB. All tests
575 were two-tailed.

576 *The motor cortex recordings*

577 The data was recorded from one female monkey (Macaca fascicularis) at 3 years of age, using a
578 10x10 microelectrode array (Blackrock Microsystems) with 0.4mm inter-electrode distance. The
579 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[53], $r_{SC} = 0.047 \pm 0.17$ (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[26]. 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,

603 presented every 300-1000 msec. For each session, all trials were either 300 msec or 500 msec long.
604 Trial length remained identical throughout a session and depended on smallest interval between
605 two tones in each session. Trials began 50 msec prior to tone burst. For spike detection, data was
606 high pass filtered with a corner frequency of 30Hz. Maximum points that were higher than 60
607 times the median of the absolute deviation from the median were classified as spikes.

608 *The Basal ganglia recordings*

609 The basal ganglia recordings are described in detail in[7]. In short, rats performed a combination of
610 a tone discrimination task and a reward-based free-choice task. Extracellular voltage was recorded
611 in the behaving rats from the NAc and VP using an electrode bundle. Spike sorting was done using
612 principal component analysis. In total, 148 NAc and 66 VP neurons across 52 sessions were used
613 for analyses (In 18 of the 70 sessions there were no neural recordings).

614 *Data Availability*

615 The data of the basal ganglia recordings is available online at https://groups.oist.jp/ncu/data and
616 was analyzed with permission from the authors. Other data are available upon request.

617 *Code Availability*

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

620 **Author Contributions**

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

622 **Acknowledgements**

623 We are extremely grateful to Oren Peles, Eilon Vaadia and Uri Werner-Reiss for providing us with
624 their motor cortex recordings, Bshara Awwad, Itai Hershenhoren, Israel Nelken for providing us
625 with their auditory cortex recordings, Kenji Doya and Makoto Ito for providing us with their basal
626 ganglia recordings, Mati Joshua, Gianluigi Mongillo and Roey Schurr for careful reading of the
627 manuscript and helpful comments and Inbal Goshen and Hanan Shteingart for discussions. This
628 work was supported by the Israel Science Foundation (Grant No. 757/16), DFG and the Gatsby
629 Charitable Foundation.

630

631 **References**

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

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

635 3. Schultz, W. Neuronal Reward and Decision Signals: from Theories to Data. *Physiol. Rev.*
636 **95,** 853–951 (2015).

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

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

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

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

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

647 9. Kim, H., Lee, D. & Jung, M. W. Signals for Previous Goal Choice Persist in the
648 Dorsomedial , but Not Dorsolateral Striatum of Rats. *J. Neurosci.* **33,** 52–63 (2013).

649 10. Ito, M. & Doya, K. Distinct Neural Representation in the Dorsolateral, Dorsomedial, and
650 Ventral Parts of the Striatum during Fixed- and Free-Choice Tasks. *J. Neurosci.* **35,** 3499–
651 3514 (2015).

652 11. Ito, M. & Doya, K. Parallel Representation of Value-Based and Finite State-Based
653 Strategies in the Ventral and Dorsal Striatum. *PLoS Comput. Biol.* **11,** 1–25 (2015).

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

657 13. Her, E. S., Huh, N., Kim, J. & Jung, M. W. Neuronal activity in dorsomedial and
658 dorsolateral striatum under the requirement for temporal credit assignment. *Sci. Rep.* **6,** 1–
659 11 (2016).

660 14. Cai, X., Kim, S. & Lee, D. Heterogeneous Coding of Temporally Discounted Values in
661 the Dorsal and Ventral Striatum during Intertemporal Choice. *Neuron* **69,** 170–182 (2011).

662 15. Kim, S., Cai, X., Hwang, J. & Lee, D. Prefrontal and striatal activity related to values of
663 objects and locations. *Front. Comput. Neurosci.* **6,** 1–13 (2012).

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

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

668 18. Fiete, I. R., Fee, M. S. & Seung, H. S. Model of Birdsong Learning Based on Gradient
669 Estimation by Dynamic Perturbation of Neural Conductances. *J. Neurophysiol.* **98,** 2038–
670 2057 (2007).

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

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

675 21. Neiman, T. & Loewenstein, Y. Covariance-Based Synaptic Plasticity in an Attractor
676 Network Model Accounts for Fast Adaptation in Free Operant Learning. *J. Neurosci.* **33,**
677 1521–1534 (2013).

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

680 23. Loewenstein, Y. & Seung, H. S. Operant matching is a generic outcome of synaptic
681 plasticity based on the covariance between reward and neural activity. *PNAS* **103,** 15224–
682 15229 (2006).

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

685 25. Loewenstein, Y. Synaptic theory of Replicator-like melioration. *Front. Comput. Neurosci.*
686 **4,** 1–12 (2010).

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

689 27. Pasquereau, B. *et al.* Shaping of Motor Responses by Incentive Values through the Basal
690 Ganglia. **27,** 1176–1183 (2007).

691 28. Day, J. J., Jones, J. L. & Carelli, R. M. Nucleus accumbens neurons encode predicted and
692 ongoing reward costs in rats. *Eur. J. Neurosci.* **33,** 308–321 (2011).

693 29. Seo, M., Lee, E. & Averbeck, B. B. Article Action Selection and Action Value in Frontal-

694 Striatal Circuits. *Neuron* **74,** 947–960 (2012).

695 30. Kim, Y. B. *et al.* Encoding of Action History in the Rat Ventral Striatum. *J Neurophysiol*
696 **98,** 3548–3556 (2007).

697 31. Stalnaker, T. A., Calhoon, G. G., Ogawa, M., Roesch, M. R. & Schoenbaum, G. Neural
698 correlates of stimulus – response and response – outcome associations in dorsolateral
699 versus dorsomedial striatum. *Front. Integr. Neurosci.* **4,** 1–18 (2010).

700 32. Guitart-Masip, M. *et al.* Go and no-go learning in reward and punishment: Interactions
701 between affect and effect. *Neuroimage* **62,** 154–166 (2012).

702 33. Fitzgerald, T. H. B., Friston, K. J. & Dolan, R. J. Action-Specific Value Signals in
703 Reward-Related Regions of the Human Brain. *J. Neurosci.* **32,** 16417–16423 (2012).

704 34. Mongillo, G., Shteingart, H. & Loewenstein, Y. The misbehavior of reinforcement
705 learning. *Proc. IEEE* **102,** 528–541 (2014).

706 35. Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of
707 distinct subpopulations of striatal neurons mimics changes in action value. *Nat. Neurosci.*
708 **15,** 1281–9 (2012).

709 36. Lee, E., Seo, M., Monte, O. D. & Averbeck, B. B. Injection of a Dopamine Type 2
710 Receptor Antagonist into the Dorsal Striatum Disrupts Choices Driven by Previous
711 Outcomes , But Not Perceptual Inference. *J. Neurosci.* **35,** 6298–6306 (2015).

712 37. Wunderlich, K., Rangel, A. & O'Doherty, J. P. Neural computations underlying action-
713 based decision making in the human brain. *PNAS* **106,** 17199–17204 (2009).

714 38. Padoa-Schioppa, C. Neurobiology of Economic choice: A Good-Based Model. *Annu. Rev.*

Neurosci. **34,** 333–359 (2011).

39. Montgomery, D. C., Peck, E. A. & Vining, G. G. *Introduction to Linear Regression Analysis*. (WILEY, 2006).

40. Gouvea, T. S. *et al.* Striatal dynamics explain duration judgments. *Elife* **4,** e11386 (2015).

41. Mello, G. B. M., Soares, S. & Paton, J. J. A Scalable Population Code for Time in the Striatum. *Curr. Biol.* **25,** 1113–1122 (2015).

42. Woolrich, M. W., Ripley, B. D., Brady, M. & Smith, S. M. Temporal autocorrelation in univariate linear modeling of FMRI data. *Neuroimage* (2001).

43. Arbabshirani, M. R. *et al.* Impact of autocorrelation on functional connectivity. *Neuroimage* (2014).

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

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

46. FitzGerald, T. H. B., Schwartenbeck, P. & Dolan, R. J. Reward-Related Activity in Ventral Striatum Is Action Contingent and Modulated by Behavioral Relevance. *J. Neurosci.* **34,** 1271–1279 (2014).

47. Vo, K., Rutledge, R. B., Chatterjee, A. & Kable, J. W. Dorsal striatum is necessary for stimulus-value but not action-value learning in humans. *Brain* **137,** 3129–3135 (2014).

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

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

737 50. Thorn, C. A., Atallah, H., Howe, M. & Graybiel, A. M. Differential Dynamics of Activity
738 Changes in Dorsolateral and Dorsomedial Striatal Loops During Learning. *Neuron* **66,**
739 781–795 (2010).

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

743 52. Ding, L. & Gold, J. I. Caudate encodes multiple computations for perceptual decisions. *J.*
744 *Neurosci.* **30,** 15747–15759 (2010).

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

Figures

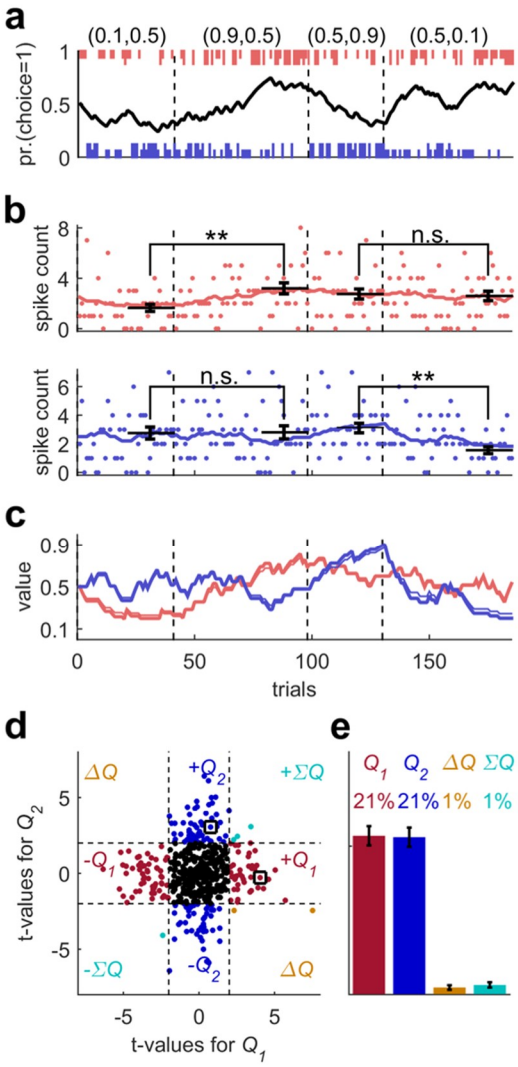

Figure 1 Model of action-value neurons **(a)** Behavior of model in example session, composed of four blocks (separated by a dashed vertical line). The probabilities of reward for choosing actions 1 and 2 are denoted by the pair of numbers above the block. Black line denotes the probability of choosing 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.

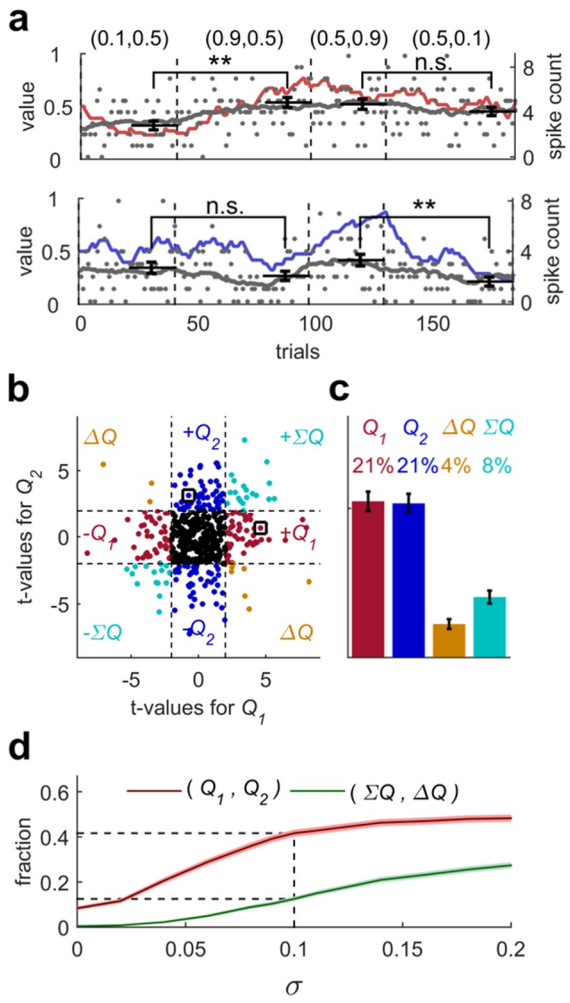

Figure 2 Erroneous detection of action-value representation in random-walk neurons **(a)** Two example random-walk neurons presented as if the sequence of spike counts of these neurons corresponds to the sequence of trials in figure 1. Gray lines and gray dots denote the neurons' firing rates and the 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.

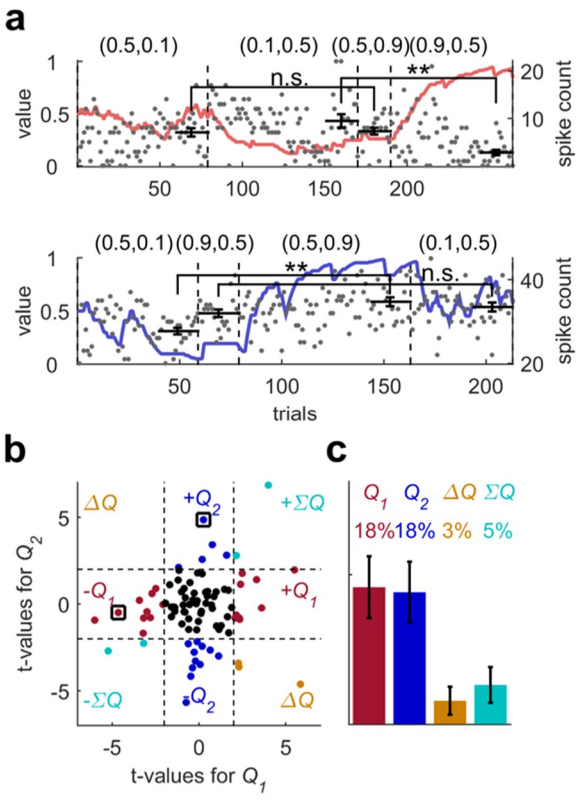

Figure 3 Erroneous detection of action-value representation in motor cortex **(a)** Two example motor cortex neurons recorded in a BMI task, presented as if the sequence of spike counts of these neurons corresponds to the sequence of trials in two sessions (one for each neuron) of operant learning used 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.

.

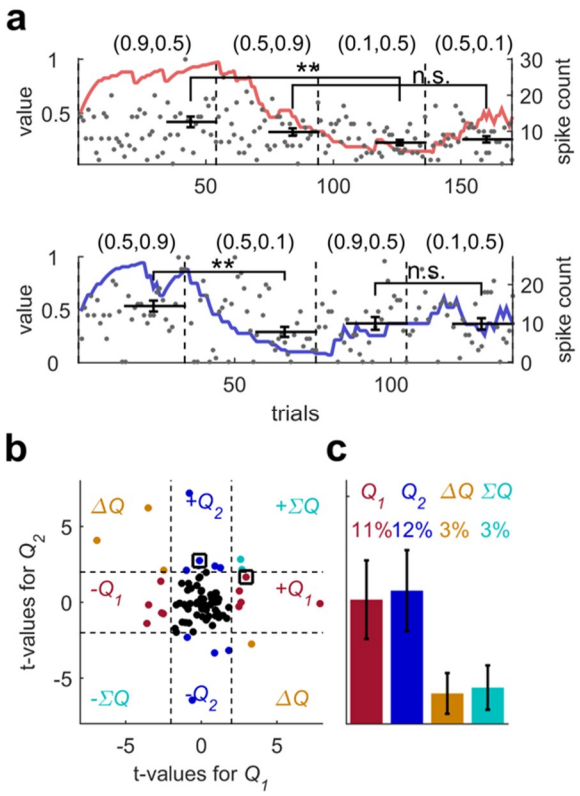

Figure 4 Erroneous detection of action-value representation in auditory cortex **(a)** Same as in Fig. 3a for two auditory cortex neurons in an anesthetized rat responding to the presentation of pure tones. **(b)** and **(c)** Population analysis. **(b)**. The t-values of 82 neurons regressed on the 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.

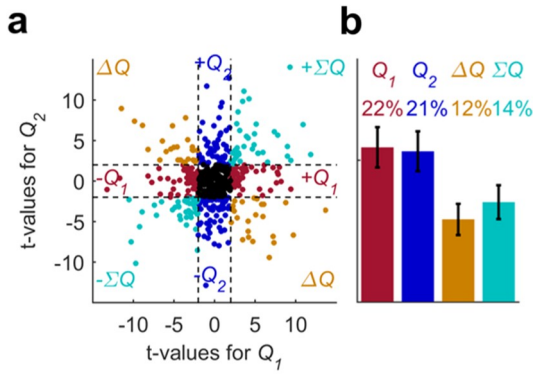

Figure 5 Erroneous detection of irrelevant action-value representation in basal ganglia **(a)** and **(b)** Population analysis. **(a)** The t-values of 214 neurons in three different phases regressed on the estimated action-values from randomly selected 642 simulated sessions (same as Fig. 4b). **(b)** 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.

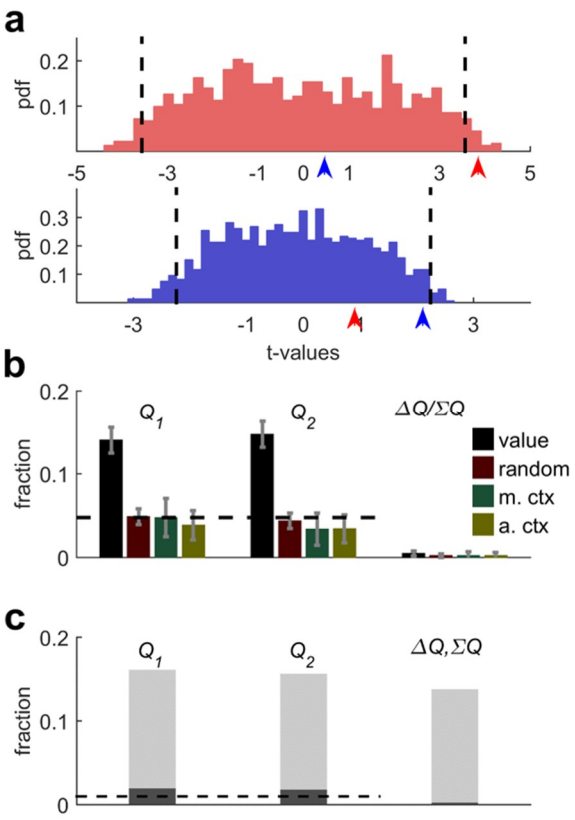

Figure 6 Permutation analysis **(a)** Red and blue figures correspond to red and blue - labeled neurons in fig. 1b, respectively. For each neuron, we computed the t-values of the regressions of its spike-count on both estimated Q-values from all sessions in Fig. 1e (excluding sessions shorter than 170 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 $p<0.01$.