nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-28983-5 OPEN

Multiplexed action-outcome representation by striatal striosome-matrix compartments detected with a mouse cost-benefit foraging task

Bernard Bloem[1,2,5,7], Rafiq Huda[2,3,6,7], Ken-ichi Amemori[4], Alex S. Abate[1,2], Gayathri Krishna[1,2], Anna L. Wilson[1,2], Cody W. Carter[1,2], Mriganka Sur[2,3] & Ann M. Graybiel[1,2]✉

Learning about positive and negative outcomes of actions is crucial for survival and underpinned by conserved circuits including the striatum. How associations between actions and outcomes are formed is not fully understood, particularly when the outcomes have mixed positive and negative features. We developed a novel foraging ('bandit') task requiring mice to maximize rewards while minimizing punishments. By 2-photon Ca^{++} imaging, we monitored activity of visually identified anterodorsal striatal striosomal and matrix neurons. We found that action-outcome associations for reward and punishment were encoded in parallel in partially overlapping populations. Single neurons could, for one action, encode outcomes of opposing valence. Striosome compartments consistently exhibited stronger representations of reinforcement outcomes than matrix, especially for high reward or punishment prediction errors. These findings demonstrate multiplexing of action-outcome contingencies by single identified striatal neurons and suggest that striosomal neurons are particularly important in action-outcome learning.

[1] McGovern Institute for Brain Research, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA. [2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA. [3] Picower Institute for Learning and Memory, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, USA. [4] Institute for the Advanced Study of Human Biology, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan. [5] Present address: Sinopia Biosciences, 600W Broadway, Suite 700, San Diego, CA 92101, USA. [6] Present address: Department of Cell Biology and Neuroscience, WM Keck Center for Collaborative Neuroscience, Rutgers University, 604 Allison Rd, Piscataway, NJ 08854, USA. [7] These authors contributed equally: Bernard Bloem, Rafiq Huda. ✉email: graybiel@mit.edu

Behavior is powerfully sculpted by learning from reinforcement, with rewards increasing and punishments decreasing the propensity to engage in specific actions[1]. The striatum has been implicated in reinforcement learning (RL) mechanisms that allow animals to adapt their behavior in changing environments by monitoring associations between actions and outcomes[2–10]. Striatal projection neurons (SPNs) encode associations between actions and rewarding outcome, i.e., outcome activity is specific for actions[7,11,12]. However, in naturalistic settings, the same action could produce both rewarding and aversive outcomes[13–17], and extensive research has shown that aversive learning also depends on the striatum[18–22]. A fundamental remaining question thus is how SPNs represent opposing reinforcement consequences of actions and use them for learning.

The underlying mechanisms of RL are often studied using probabilistic and non-stationary bandit tasks, which require trial-and-error learning in order to maximize only rewards to be obtained[1,23–32]. We developed a 'cost–benefit bandit (CBB) task' to bring the advantages of bandit task protocols to address this mixed outcome context. In this dynamic foraging task, each of two available actions is probabilistically linked to outcomes of opposing valence, and the mice learn action–outcome contingencies to maximize reward delivery and minimize air puff delivery.

To address the question of whether the striatal system learns about the rewarding and aversive outcomes of actions in parallel, or whether it learns the overall value of actions, we recorded the activity of thousands of SPNs during the CBB task by 2-photon Ca^{++} imaging and tracked their encoding of action–reward–airpuff associations in the mid-dorsal sector of the caudoputamen. We found large numbers of SPNs responding in relation to both rewarding and aversive outcomes of a given action. To our surprise, among all outcome-responsive SPNs, equal numbers of neurons were active in relation to outcomes with the same or opposite valence. On a population level, both outcomes were represented independently in two partially overlapping populations of neurons, resulting in a multiplexed representation. This finding is important, as it suggests that striatal reward-related activity might not reflect the integrated value, but rather, specific outcomes.

The RL perspective hypothesizes that the value of actions is updated using prediction errors (PEs). The striatum has been implicated in the PE signals[33,34], which are transferred to dopamine-related circuitry[35–39]. Yet it has been unclear whether SPNs represent PEs by integrating the outcomes of opposing valence, or whether they independently represent reward prediction errors (RPEs) and punishment prediction errors (PPEs). Our findings suggest that the activity of outcome-selective SPNs is better accounted for by separate RPE and PPE as compared to single integrated PEs, underscoring parallel coding of rewarding and aversive learning variables.

Computational work has suggested that SPNs in the striosome compartment might specifically function in shaping PE signals, thereby providing learning signals for matrix neurons to learn action values[40–44]. It is thus possible that PE-based updating, and modulation of dopaminergic activity, could arise, in part at least, from striosomes, based on anatomical and electrophysiological studies demonstrating strong projections from the striosome compartment to dopamine-containing neurons in the substantia nigra[45–49].

In parallel, the use of approach-avoidance paradigms has led to the suggestion that striosomal circuits could be part of a conserved mechanism for facing critical decisions requiring an estimate of cost-benefit evaluation[14,16,50–52]. We addressed this issue by visually identifying striosomal SPNs (sSPNs) and matrix SPNs (mSPNs) in and around the most dorsal band of striosomes by using their birthdates to label striosomes[48,53,54]. sSPNs and mSPNs exhibited marked differences in their encoding of outcomes. Outcome encoding by sSPNs was particularly strong when RPE or PPE was high, a finding in accord with the proposed role for striosomes as being part of the circuit that calculates the learning signals[41]. Notably, we did not find differential encoding of motor behavior. Hence, the learning function may be a principal function of striosomes added to other shared attributes with neurons of the surrounding matrix[55].

We suggest that multiplexed encoding of action–outcome associations and PEs are key characteristics of large numbers of SPNs in the anterodorsal striatum, that striosomes in this region are particularly sensitive to error signals of both positive or negative valence outcomes, and that models incorporating these features could be of great value in understanding how striatal circuits underpin adaptive behaviors.

Results

A dynamic foraging task with rewarding and aversive outcomes. Mice were trained on the CBB task with rapidly changing action–reward and action–air puff contingencies. Head-fixed mice, with their forepaws on a wheel (Fig. 1a), initiated trials by holding the wheel still for 2 s (Fig. 1b). Mice had 3 s to make a leftward or rightward response. Both actions were probabilistically linked to receipt of rewarding (water) and/or punishment (air puff delivered to the face) outcomes. The action–reward and action–puff contingencies changed rapidly without cueing, and the reward and puff block changes were made independent of one another (Fig. 1c) after mice received 6–15 rewards or avoided this number of puffs. Reward and puff action–outcome contingencies switched independently of one another; actions could lead to one of four outcome pairs (reward–puff, reward–no puff, no reward–puff, and no reward–no puff). Mice performed hundreds of trials (average: 366; range: 247–585) per session, with tens of reward (mean: 20.4; range: 14–34) and puff (mean: 22.9; range: 14–38) block switches. They reliably adapted their behavior within blocks with a given action–reward or action–puff contingency (Fig. 1d) and around action–reward or action–puff contingency switches (Fig. 1e). We compared choice behavior as a function of the actions and outcomes of the previous two trials (Fig. 1f), and we performed an autoregression analysis using the last five trials (Fig. 1g). Both analyses show that mice incorporate outcomes of multiple past trials rather than relying on a strict win-stay/lose-shift strategy.

SPN activity represents associations between actions and multiple outcomes. We used 2-photon imaging to measure Ca^{++} activity of GCaMP6s-expressing SPNs via a cannula placed above the left striatum (Fig. 2a). In every session, we recorded a new field of view, resulting in a total of 5831 unique neurons (75 sessions; 13 mice). Every field of view contained SPNs of both striosome ($n = 2249$) and matrix ($n = 3582$) compartments (Fig. 2b). Only low percentages of striosomal neurons were labeled, but with this preparation we could identify the compartments based on labeling of the neuropil. We regarded every neuron in a red-labeled neuropil cluster as being a striosomal SPN. There was no significant effect of the imaging cannula on the number of sessions required to learn this task (Supplementary Fig. 1a). In trained mice, the response time was higher in mice with a cannula ($p < 0.01$; independent t-test), but otherwise the performance of the mice was identical (Supplementary Fig. 1b–e).

We identified individual Ca^{++} events in $\Delta F/F$ traces using a custom algorithm (see the "Methods" section, Supplementary Fig. 2a). By use of a chi-square test, we compared the number of trials with a Ca^{++} transient in the 3 s following outcome delivery to identify neurons with selective responses to specific actions and outcomes. We did not observe neurons that were tonically active but consistently inhibited by one or more task event. Many neurons exhibited strong activity selective for rewarding outcomes,

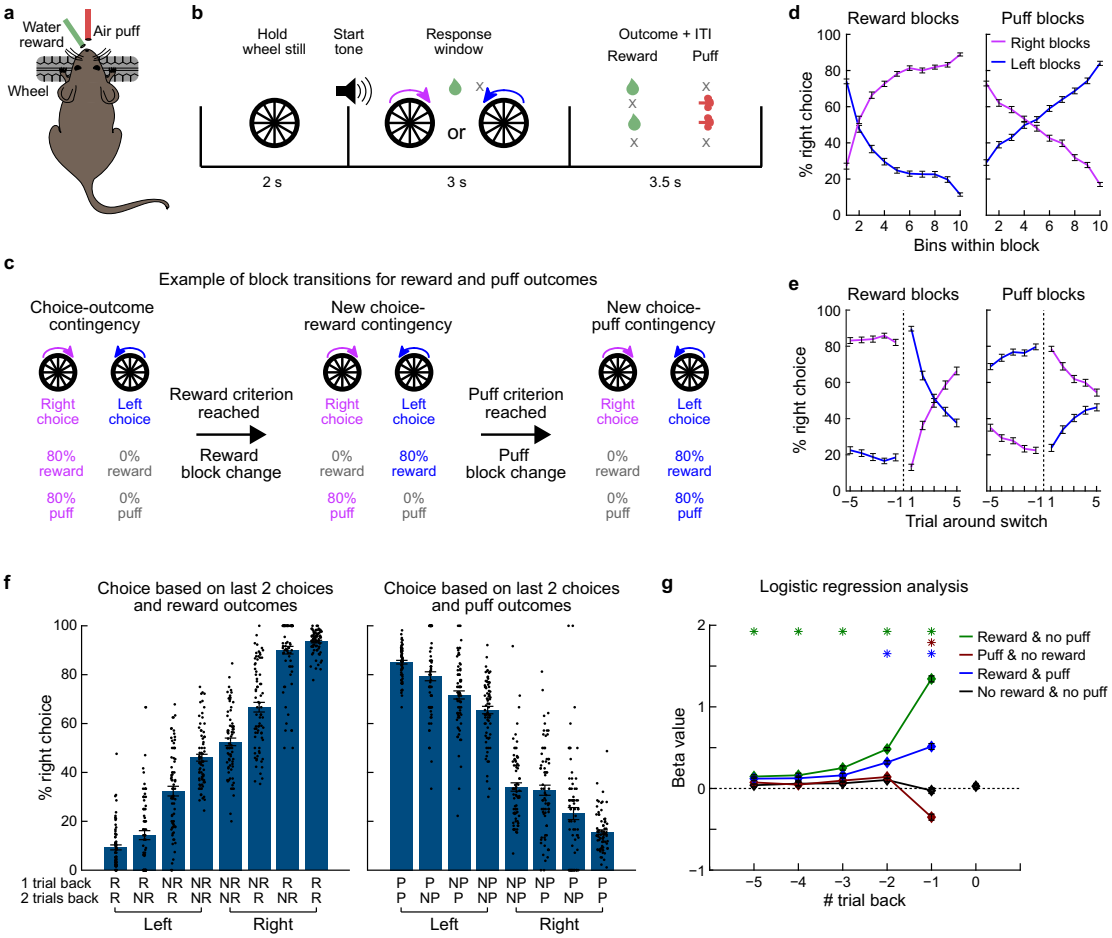


Fig. 1 The cost-benefit bandit task. a Behavioral setup. Head-fixed mice reported choices by rotating a wheel to the left or the right. Water rewards and air puffs were delivered via two tubes. **b** Trial procedure. Mice initiated trials by holding the wheel still for 2 s. An auditory go-cue signaled trial start, after which the mouse had 3 s to move the wheel 15° in either direction. Left and right actions were probabilistically linked to reward and air puff delivery, giving four possible outcome combinations: reward-no puff, puff-no reward, reward-puff, and no reward-no puff. **c** Action–reward and action–puff contingencies changed in block of trials independently of each other. **d** Choices within the reward and puff blocks. Trials within blocks were grouped into 10 bins. Error bars show SEM across sessions ($n = 75$ sessions from 13 mice). **e** Behavioral adaption during five trials before and after reward and puff block switches (mean ± SEM, $n = 75$ sessions). **f** Mouse decisions were affected by the outcomes and actions in the last two trials (mean ± SEM, $n = 75$ sessions). For simplicity, only trials in which the mouse made the same action in the last two trials were included (R, reward; NR, no reward; P, puff; NP, no puff). Trials in which the mouse did not make the same response in the last two trials were omitted. **g** Regression coefficients (mean ± SEM) for the 4 types of outcomes (reward/no puff, puff/no reward, reward/puff, and no reward/no puff) for the preceding five trials, as calculated using an auto-regressive model (reward, no puff: −1: $p = 1e{-}47$, −2: $p = 1e{-}19$, −3: $p = 1e{-}6$, −4: $p = 0.0065$, −5: $p = 0.0018$; puff, no reward: −1: $p = 1e{-}8$; reward and puff: −1: $p = 1e{-}17$, −2: $p = 1e{-}6$; $n = 13$ mice, two-sided t-test comparing the beta values of reward, puff and both outcomes with no outcome, $^*p < 0.05$). Source data are provided as a Source Data file.

puff outcomes and chosen actions (Fig. 2c, d; examples of neurons in Supplementary Fig. 2b, c). More neurons responded to the absence than to the presence of rewarding outcome ($p < 0.05$, $t = 2.89$, df = 12, paired t-test) and more neurons responded to the presence than to the absence of aversive puff outcome ($p < 0.001$, $t = 6.34$, df = 12, paired t-test). We found no significant difference in the number of neurons active for actions ipsilateral or contralateral to the imaged hemisphere.

Following evidence that SPNs encode action–outcome associations for reward[11,12], we tested whether reward– and puff–outcome activity was selective for specific actions with two complementary approaches. First, for every neuron, we used a stepwise logistic regression analysis to determine which factor(s) among the chosen action and received outcomes could best predict whether or not a transient occurred in a trial (using a cutoff of $p < 0.05$). We then identified neurons active in relation to a factor as neurons that included that factor in their model. The activity of the neurons identified in this analysis, averaged over all trials for a given action–outcome combination, is illustrated in Fig. 2e. The same analysis using half the trials for model fitting and the other half for calculating the average responses produced similar results (Supplementary Fig. 2d). We also used a more conservative approach, employing sequential chi-square analyses. We found many reward-selective and puff-selective SPNs among the action-selective SPNs (Supplementary Fig. 2e, f) and action selective neurons among reward- and puff-selective neurons (Supplementary Fig. 2g, h). Finally, we used a regression analysis to find the regression coefficients for actions, reward and puff outcomes for every neuron imaged (Supplementary Fig. 2i, j). These findings indicate that the striatal neurons encoded action–puff contingencies similarly to their encoding of action–reward contingencies.

We asked whether action–outcome contingency-related activity was present in the first trial after a block switch, or instead developed within a

block with a given action–outcome contingency. For neurons with significant encoding of action–outcome contingencies, we quantified the percentage of trials with these actions and outcomes in which a Ca^{++} transient occurred. Neurons that were active specifically for an action-reward combination were more active in the first such trial after a switch as compared to other trials with the same action and outcome (Supplementary Fig. 2k, $p < 0.005$, $t = 3.51$, df = 12, repeated measures t-test), as were neurons active specifically for an action–puff combination ($p < 0.001$, $t = 6.66$, df = 12). No difference was observed for neurons that were active for a given action and the omission of reward or puff.

Next, we asked whether action–reward and action–puff contingencies were represented by separate or overlapping populations of SPNs. We found in our stepwise regression analysis that 34.66 ± 1.1% of neurons (mean ± SEM, $n = 13$ mice) encoded combinations of actions and both rewards and puffs (Fig. 2f). We used a second regression analysis to identify the neurons that encoded each of the 8 possible action–reward–puff associations (Fig. 2g; examples of neurons in Supplementary Fig. 2l). The average activity of action–reward–puff neurons for the 8 possible trial types (chosen action × reward outcome × puff outcome) showed selective responses for specific combinations (Fig. 2h), which we again confirmed in an additional analysis in which half of the trials were used for identifying neuron-types and the other half for quantifying the average responses

Fig. 2 Representations of action-reward-puff combinations by SPNs. a Imaging setup and preparation. **b** Example of imaging field-of-view showing GCaMP6s-expressing neurons (green). tdTomato (red) expressed in the cell bodies and neuropil demarcates the striosome compartment. Seventy-five of such fields-of-view were imaged in a total of 13 mice. **c** Two sample neurons during 10 min of recording. Neuronal transients occurred selectively for action–reward–puff associations during the outcome period. **d** Percentage (mean ± SEM) of neurons selectively responding to chosen action (left: 12.1 ± 1.4%, right: 7.8 ± 1.2%), reward outcomes (reward: 13.1 ± 1.8%, no reward: 22.7 ± 2.2%), and puff outcomes (puff: 23.0 ± 2.5%, no puff: 4.8 ± 0.6%). More neurons responded to the absence than to the presence of rewarding outcome ($p = 0.013$, $t = 2.89$, $n = 13$ mice, two-sided paired t-test), and more neurons responded to the presence than to the absence of aversive puff outcome ($p = 0.00037$, $t = 6.34$, $n = 13$, two-sided paired t-test). $^{*}p < 0.05$; $^{***}p < 0.001$; two-sided paired t-test. **e** Activity of neurons that represented action-outcome associations, averaged over all trials for a given action-outcome combination per session. Activity of each neuron was normalized to its average transient rate for all trials (left: action–reward representations; right: action–puff representations). **f** Percentage of neurons per mouse with chosen action, reward, puff, 2-way interactions, and 3-way interactions as significant predictors in a stepwise regression analysis (mean ± SEM, $n = 13$). **g** Percentage of neurons with activity selective for the 8 possible action–reward–puff combinations, identified in a stepwise regression analysis (mean ± SEM, $n = 13$). **h** Average activity of the neurons in (**g**), shown separately for the 8 different trial types. Neuronal activity was normalized to the average transient rate. **i** Joint distribution of the number of neurons responding to different action–reward and action–puff associations. **j** Among the neurons that selectively responded to both an action–reward and action–puff associations, the vast majority had a stable action-preference ($p = 1e{-}7$, two-sided paired t-test, $n = 13$, mean ± SEM). Source data are provided as a Source Data file.

(Supplementary Fig. 2m). Neurons encoding all three factors had stable action encoding. Among the neurons that encoded both an action–reward and action–puff association, almost all encoded both outcomes for the same action (Fig. 2i, j; same action: 8.5 ± 0.9% of all neurons, different: 0.1 ± 0.1%; $n = 13$ mice, $p < 0.001$, paired t-test, $t = 10.35$, $df = 12$). These analyses demonstrated that SPNs can encode associations of multiple outcomes with a given action.

Multiplexed SPN encoding of reward and aversive outcomes of actions. The responses of the individual SPNs could represent outcome value (i.e., value coding) if the activity signaled good (reward and/or no puff) or bad outcomes (no reward and/or puff). Alternatively, action–outcome encoding could be multiplexed such that, as a population, the SPNs would respond to all combinations of actions and outcomes without being biased toward specific combinations. In this case, partially overlapping populations of single SPNs would encode each action–outcome contingency, and single SPNs encoding both outcomes would not necessarily represent the combined value of the outcomes. Our results favor this latter form of multiplexed action–outcome encoding (Fig. 3).

Many single SPNs responded to both reward and puff or to the absence of both. Chi-square analyses detected puff-selective and no-puff selective SPNs among the reward neurons and no-reward neurons (Fig. 3a). With ANOVA, we found no interaction in the distribution of puff/no-puff neurons amongst the reward/no-reward population (Fig. 3a; $p = 0.83$, $F = 0.05$, $n = 13$ mice), and only a significant main effect of puff outcome ($p < 0.001$, $F = 33.74$). Conversely, when we compared the distribution of reward and no-reward selective SPNs among the puff- and no-puff-selective neurons (Fig. 3b) with ANOVA, we found a significant main effect of reward outcome ($p < 0.05$, $F = 4.28$) but no interaction in the distribution of reward/no-reward neurons amongst the puff/no-puff neuron population (Fig. 3b; $p = 0.84$, $F = 0.04$). These findings suggest that the selectivity of a given SPN for one outcome did not depend on the selectivity for the other outcome. The activity of all neurons responding to both outcomes is shown in Fig. 3c, averaged over all trials for each of the four different reward–puff combinations.

To compare value encoding and multiplexed outcome encoding directly, we identified 'value neurons', i.e., SPNs whose activity reflected that an action was good or was bad (good outcomes: reward and no puff; or bad outcomes: puff and no reward), and 'non-value neurons', i.e., SPNs that responded in relation to the presence or absence of both outcomes. We found similar proportions of value and non-value neurons among the recorded neurons; $p > 0.05$, $t = 0.4$, $df = 12$, paired t-test, Fig. 3d). The value and non-value neurons did not detectably differ in their selectivity for reward or puff (Fig. 3e, Supplementary Fig. 3a).

The representation of both 'good' and 'bad' outcomes in single neurons could potentially be accounted for by differences in action selectivity for both outcomes (a neuron could be active in trials where left choices lead to reward or right choices lead to air puffs). This possibility appeared unlikely, because we found stable action selectivity (Fig. 2j). We further tested this idea by quantifying the activation of neurons encoding action–reward–-puff interactions in all eight action–reward–puff trial types. SPNs that encoded the good or bad outcome of one action, as a population, were not activated when the mouse selected the opposite action and received outcomes of opposite value (Supplementary Fig. 3b). Hence, the population activity of the striatal SPNs encoded combinations of outcomes in a multiplexed manner so that their activity reflected multiple outcomes of actions rather than the overall value of an action.

Encoding of multiplexed prediction errors. We used RL models to gain more insight into the behavior and to test how trial-by-trial PEs for rewarding and aversive outcomes were represented in the SPNs. In such models, costs and rewards associated with actions are typically combined into one scalar value, and for every action one action-value is learned. However, the observed activity of SPNs in mice performing the CBB task suggested that these outcomes could be represented in parallel for cost-based and reward-based associations. We therefore adapted existing RL models to integrate costs and benefits by means of two alternative approaches. In the first model, one set of action values (Q) is used to model the expected overall value of the two possible actions. When outcomes are delivered, reward and puff outcomes are weighted with the sensitivity parameters β_{rew} and β_{puff} and combined into one outcome value V (Eq. (1)), which is compared with the predicted value Q to calculate one PE (Eq. (2)), and to update the single set of Q-values (Eq. (3)). Three learning rates were used (for chosen action: α_1: if outcome was delivered; α_2: if outcome was not delivered; for non-chosen actions: γ).

$$V(t) = \beta_{rew} * \mathrm{Reward}(t) + \beta_{puff} * \mathrm{Puff}(t) \quad (1)$$

$$\mathrm{PE}(t) = V(t) - Q^{chosen}(t) \quad (2)$$

$$Q^{chosen}(t+1) = \begin{cases} Q^{chosen}(t) + \alpha_1 * \mathrm{PE} & \text{if reward/puff delivered} \\ Q^{chosen}(t) + \alpha_2 * \mathrm{PE} & \text{if no outcome delivered} \end{cases} \quad (3)$$
$$Q^{unchosen}(t+1) = Q^{unchosen}(t) * (1-\gamma)$$

This single set of action values, which represents the value combining the reward and puff outcomes, is used to make a

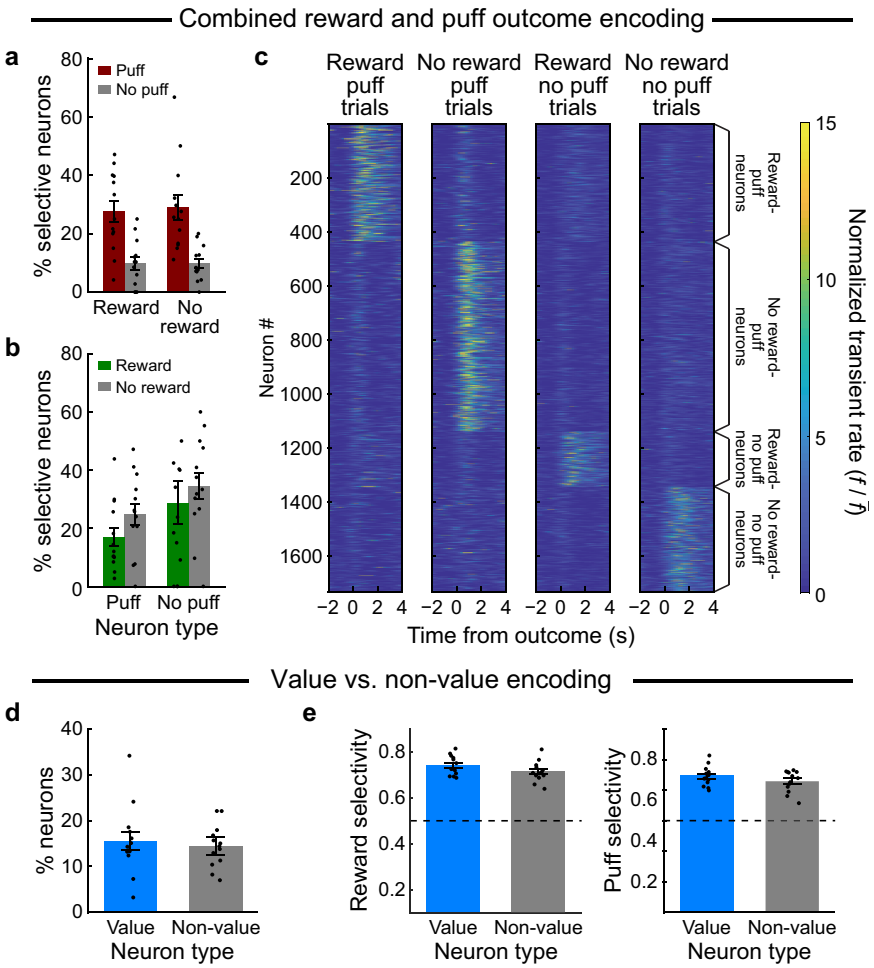


Fig. 3 Combined reward-puff representations in single neurons do not always reflect outcome value. a Percentage (mean ± SEM) of reward-selective (puff: 27.6 ± 3.6%; no puff: 9.7 ± 2.2%) and no-reward-selective (puff: 29.0 ± 4.2%; no puff: 9.7 ± 1.6%) neurons that were selective for the puff outcome, as identified using chi-square analysis ($n = 13$ mice for all panels, average percentage per mouse is shown). There was a significant main effect of the percentage of puff neurons ($p = 1e{-}7$, repeated measures ANOVA) but no reward-puff interaction ($p = 0.83$). **b** Percentage of puff- and no-puff-selective neurons that were selective for reward/no-reward outcome (puff selective: reward: 16.9 ± 3.2%, no reward: 24.8 ± 3.7%; no-puff-selective: reward: 28.7 ± 7.4%, no reward: 34.5 ± 4.6%, mean ± SEM, $n = 13$). There was a significant main effect of puff neurons ($p = 0.044$, repeated measures ANOVA) but no interaction between reward and puff ($p = 0.84$). **c** A stepwise regression analysis identified neurons with reward and puff outcome interactions. Session-averaged activity of these neurons is shown for trials with different reward-puff combinations. **d** Neurons were split into value and non-value type based on whether they responded to reward and puff oppositely or not. The proportion of neurons classified to these two types was not different (value: 15.6 ± 2%, non-value: 14.5 ± 2%, mean ± SEM percentage across 13 mice). **e** The selectivity of value and non-value neurons in (**d**) is not significantly different for reward (left panel; value: 0.74 ± 0.01 (mean ± SEM), non-value: 0.71 ± 0.01, $p > 0.05$, $t = 2.1$, df = 12, two-sided paired t-test, $n = 13$) or puff (right panel; value: 0.69 ± 0.01, non-value: 0.67 ± 0.01, $p > 0.05$; $t = 1.03$, df = 12, two-sided paired t-test, $n = 13$). Source data are provided as a Source Data file.

decision (Eq. (4)).

$$P^R(t) = \frac{1}{1+e^{\beta*(Q^R(t)-Q^L(t))}}$$
$$P^L(t) = 1 - P^R(t) \quad (4)$$

We refer to this model as the *integrated model*.

In the second model (Fig. 4), two parallel sets of Q-values are estimated: Q_{rew} for reward and Q_{puff} for puff. PEs are calculated separately for both outcomes (Eq. (5)), and both sets of Q-values are updated in parallel using two sets of learning rates (α and γ, respectively), for reward and for puff (Eq. (6)).

$$\text{RPE}(t) = \text{Reward} - Q_{\text{rew}}^{\text{chosen}}(t)$$
$$\text{PPE}(t) = \text{Puff} - Q_{\text{puff}}^{\text{chosen}}(t) \quad (5)$$

$$Q_{\text{rew}}^{\text{chosen}}(t+1) = \begin{cases} Q_{\text{rew}}^{\text{chosen}}(t) + \alpha_{\text{rew}} * \text{RPE} & \text{if rewarded} \\ Q_{\text{rew}}^{\text{chosen}}(t) + \alpha_{\text{unrew}} * \text{RPE} & \text{if not rewarded} \end{cases}$$
$$Q_{\text{rew}}^{\text{unchosen}}(t+1) = Q_{\text{rew}}^{\text{unchosen}}(t) * (1 - \gamma_{\text{rew}})$$
$$Q_{\text{puff}}^{\text{chosen}}(t+1) = \begin{cases} Q_{\text{puff}}^{\text{chosen}}(t) + \alpha_{\text{puff}} * \text{PPE} & \text{if puffed} \\ Q_{\text{puff}}^{\text{chosen}}(t) + \alpha_{\text{nopuff}} * \text{PPE} & \text{if not puffed} \end{cases}$$
$$Q_{\text{puff}}^{\text{unchosen}}(t+1) = Q_{\text{puff}}^{\text{unchosen}}(t) * (1 - \gamma_{\text{puff}}) \quad (6)$$

Decisions are then based on the integration of both sets of Q-values (Eq. (7)), weighted by β_{rew} and β_{puff} to allow for differences

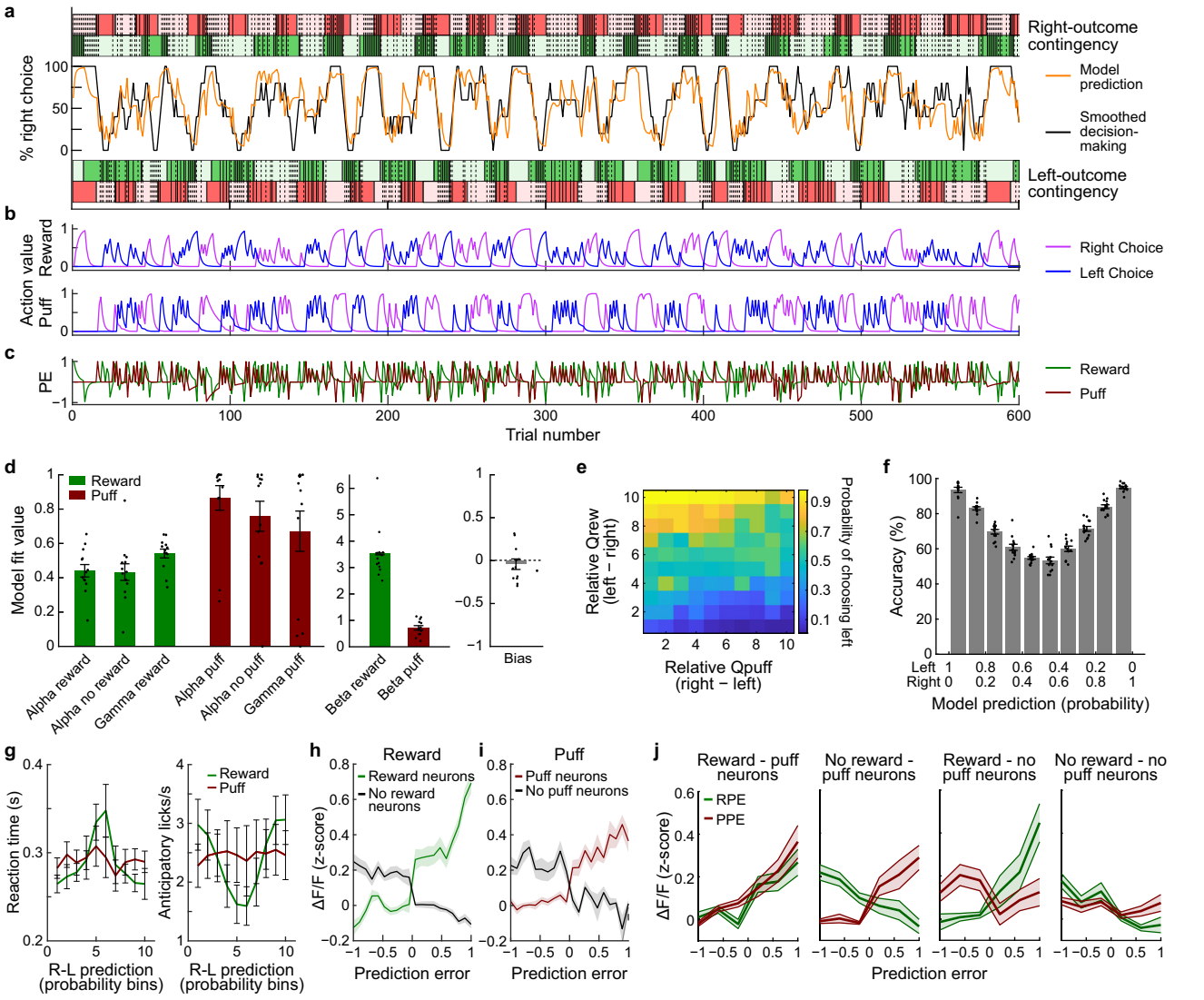


Fig. 4 Outcome neurons are modulated by RPE and PPE. a Session example. Action–outcome contingency is shown as colored blocks (reward: green; puff: red) for right (top) and left (bottom) choices. Dark and light red/green shading indicates, respectively, 80% and 0% probability of outcome delivery. Within blocks, solid lines indicate trials in which the outcome was delivered, and dotted lines indicate trials in which the outcome was not delivered. Middle lines show the smoothed decisions by mice (black) and model prediction (orange). **b** Estimated *Q*-values for the reward (top) and puff (bottom) for both actions for every trial. **c** RPE and PPE calculated for every trial by subtracting the *Q*-value of the chosen action from the outcome. **d** Average model parameters estimated by the RL model (alpha: learning rate for trials with reward/no reward and puff/no puff; gamma: outcome specific forgetting rate; beta: inverse temperature; bias: directional bias) (mean ± SEM, $n = 13$ mice). **e** Fraction of trials in which the mouse chose the left action as a function of relative reward Q-values and relative puff Q-values. **f** Accuracy as a function of the probability of left and right actions as predicted by the model (mean ± SEM, $n = 13$). **g** Trial-by-trial model predictions correlated with reaction time (left) and anticipatory licking (right; $n = 13$ mice, data are mean ± SEM). **h** $\Delta F/F$ response during outcome period was correlated with RPE in reward neurons (green, $r = 0.89$; $p = 1e{-}8$, two-sided Pearson correlation) and in no-reward neurons (black, $r = -0.76$; $p = 0.026$). Data shown represent mean ± SEM. **i** $\Delta F/F$ was significantly correlated with PPE in puff neurons (red, $r = 0.92$; $p = 0.0008$, two-sided Pearson correlation), but not in no-puff neurons (black, n.s.). Data shown represent mean ± SEM. **j** Activity of neurons that responded to both reward and puff outcomes (Fig. 3c) is modulated by RPE and PPE (reward-puff neurons: RPE: $r = 0.68$, $p = 0.0051$; PPE: $r = 0.68$, $p = 0.0052$; no-reward-puff neurons: RPE: $r = -0.82$, $p = 0.00017$; PPE: $r = 0.45$, $p = 0.009$; reward-no-puff neurons: RPE: $r = 0.88$, $p = 0.000016$; PPE: $r = -0.11$, $p = 0.023$; no-reward-no-puff neurons: RPE: $r = -0.40$, $p = 0.012$; PPE: $r = -0.5$; $p = 0.050$, two-sided Pearson correlation). Data shown represent mean ± SEM. Source data are provided as a Source Data file.

in sensitivity to reward and puff outcomes.

$$\begin{aligned}\Delta Q^{R-L}_{\mathrm{rew}}(t) &= \beta_{\mathrm{rew}} * (Q^{R}_{\mathrm{rew}}(t) - Q^{L}_{\mathrm{rew}}(t))\\ \Delta Q^{R-L}_{\mathrm{puff}}(t) &= \beta_{\mathrm{puff}} * (Q^{R}_{\mathrm{puff}}(t) - Q^{L}_{\mathrm{puff}}(t))\\ P^{R}(t) &= \frac{1}{1+\mathrm{e}^{(\beta_0 - \Delta Q^{R-L}_{\mathrm{rew}}(t) + \Delta Q^{R-L}_{\mathrm{puff}}(t))}}\\ P^{L}(t) &= 1 - P^{R}(t)\end{aligned} \quad (7)$$

We refer to this model as the *parallel model*.

The parallel model predicted behavior slightly but non-significantly better in cross-validation (Supplementary Fig. 4a). Models with reduced number of parameters had lower prediction accuracy in a test set (Supplementary Fig. 4b). We also tested the history dependency of decision-making by setting the learning rates to 1, resulting in a win-stay/lose-shift model, and again found that this significantly impaired model performance.

Next we tested whether the PE variables from the two models could account for the neuronal activity using a stepwise logistic regression model. We found that both reward and puff outcomes

could account for firing in more neurons than could the integrated value of the outcome (Supplementary Fig. 4c; reward vs. combined: $p < 0.05$, $t = 2.99$; puff vs. combined: $p < 0.05$, $t = 2.62$; df = 12, paired t-tests). Similarly, the addition of separate RPE and PPE significantly improved the prediction of transients in more neurons than the addition of an integrated PE (RPE vs. combined PE: $p < 0.05$, $t = 2.27$; PPE vs. combined PE: $p < 0.05$, $t = 2.81$; df = 12, paired t-tests). We confirmed these results using a partial regression analysis performed on the residuals of the activity against the outcomes, against the residuals of the prediction errors against the outcomes. There were significantly more neurons with a significant effect of adding RPE or PPE than with a combined PE (Supplementary Fig. 4d, reward: $p < 0.005$, $t = 6.12$, df = 12; puff: $p < 0.001$, $t = 6.84$, df = 12, repeated measures t-test).

We further analyzed these results by focusing on 'value neurons', whose activity reflected whether the outcome was good or bad, and 'non-value neurons', whose activity reflected a mix of good and bad outcomes (Supplementary Fig. 4e). In 'non-value neurons', we found significantly more neurons whose firing could be accounted for by puff and reward outcomes than by combined outcomes (reward: $p < 0.01$, $t = 3.13$, df = 12; puff: $p < 0.005$, $t = 3.72$, df = 12; repeated measures t-test), and significantly more neurons with firing that could be explained by PPE than by combined prediction errors ($p < 0.005$, $t = 3.46$, df = 12). For 'value neurons', there was a similar trend but without statistical significance; but the parallel model still accurately described neuronal firing, even in neurons that conform to an integrated outcome encoding scheme. Therefore, we used the parallel cost-benefit RL model to derive reward- and puff-specific action values and prediction errors (Fig. 4a–d).

With this model, we found that the relative difference in the inferred positive and negative action values predicted the selected actions of the mice (Fig. 4e, f). The model predictions were systematically correlated with reaction times and anticipatory licking in the 1-s period preceding the outcome (Fig. 4g). Hence, the model predictions could be generalized to behavioral variables beyond the choices and their outcomes that were used to fit the model.

We further characterized the modulation of the observed reward and puff outcome activity by RPE and/or PPE. The activity of reward-responsive and puff-responsive neurons was correlated with model-derived RPE (reward neurons: $r = 0.89$, $p < 0.001$; no-reward neurons: $r = -0.76$, $p < 0.05$; Fig. 4h) and PPE (puff neurons: $r = 0.92$, $p < 0.001$; no-puff neurons: $r = -0.51$, $p > 0.05$; Fig. 4i) during the 3 s window after outcome delivery. RPE encoding was significantly stronger in 'reward neurons' than in 'puff neurons' ($p < 0.001$, $z = 3.37$, Fisher z transformation) and stronger in 'no-reward neurons' than in 'no-puff neurons' ($p < 0.005$, $z = -2.73$). PPE encoding was stronger in 'puff neurons' than in 'reward neurons' ($p < 0.05$; $z = 2.03$) and stronger in 'no-puff neurons' than in 'no-reward neurons' ($p < 0.001$, $z = -3.49$). In neurons modulated by both reward and puff outcomes, as found by use of stepwise regression analysis (Fig. 2), outcome-related activity was modulated by both RPE and PPE (reward–puff neurons, RPE: $r = 0.68$, $p < 0.01$; PPE: $r = 0.68$, $p < 0.01$; no-reward–puff neurons, RPE: $r = -0.82$; $p < 0.001$; PPE: $r = 0.45$; $p < 0.05$; reward–no-puff neurons, RPE: $r = 0.88$; $p < 0.001$; PPE: $r = -0.11$; $p < 0.05$; no-reward–no-puff neurons: RPE: $r = -0.40$; $p > 0.05$; PPE: $r = -0.5$; $p < 0.05$; Fig. 4j). The modulation of neuronal activity was observed in the period after the outcome was delivered. An alternative interpretation, that this activity reflected differences in reward and puff anticipation before outcome delivery, seems unlikely, given that it should produce higher neuronal activity in trials with high reward/puff expectation, the opposite of what we observed.

Stronger encoding of outcomes and PEs in striosomes than in nearby matrix. We tested the hypothesis that mSPNs differentially encode 'motor' and sSPNs 'limbic' reinforcement-related information. First, we identified action- and outcome-encoding neurons using chi-square analysis (Fig. 5a). We did not find a compartmental difference between the number of neurons encoding chosen actions. However, more sSPNs than mSPNs were selectively activated for reward ($p < 0.05$, $t = 2.30$, df = 12, repeated measures t-test), puff ($p < 0.05$, $t = 2.23$, df = 12) or no-puff ($p < 0.05$, $t = 2.77$, df = 12) outcomes. The distribution of single-neuron regression coefficients concurred with these results (Fig. 5b); we observed significant differences in outcome encoding between the striosomal and matrix populations (reward: $p < 0.005$, KS = 0.053; puff: $p < 0.01$, KS = 0.051; $n = 2249$ sSPNs and 3582 mSPNs, Kolmogorov–Smirnov test), but not in action encoding.

Task-related actions are strongly linked to outcomes. We therefore tested whether mSPNs had a stronger encoding of movements than sSPNs during intertrial intervals (ITIs), times during which the actions were likely to be less related to task performance. We identified the onset and offset of wheel movement bouts during ITIs. Similar proportions of neurons in each compartment were modulated by movement acceleration and deceleration regardless of movement direction (Supplementary Fig. 5a–f). Similar proportions of sSPNs and mSPNs were modulated for maximum acceleration or deceleration of wheel movements during ITIs (Fig. 5c). We similarly aligned neuronal activity to the onset of licking bouts during the entire session or ITI period and found no differences in the proportion of responsive sSPNs and mSPNs (Supplementary Fig. 5g, h). Finally, we aligned licks during ITIs to the time of neuronal events to quantify the number of neurons with activity coincident with licking (Fig. 5d) and again found no difference. These results point to a similar activation of sSPNs and mSPNs in relation to movements during trials and ITIs.

Striosomes have been proposed to function as critics in actor-critic RL models[41]. Our finding that sSPNs had a stronger encoding of both positive and negative outcomes is consistent with this view. We further tested this proposition by comparing modulation of sSPN and mSPN activity by RPE and PPE. We used regression analysis to identify neurons for which activity could be better accounted for if RPE and/or PPE were included as regressors. The percentage of sSPNs was higher for both RPE and PPE (Fig. 5e, Supplementary Fig. 5i). Moreover, reward- and puff-responsive neurons in striosomes were more strongly modulated by, respectively, RPE and PPE than those in the matrix (Fig. 5f). As a result, the relative sSPN-to-mSPN activation was correlated with RPE (reward neurons: $r = 0.77$, $p < 0.01$; no-reward neurons: $r = -0.62$, $p = 0.06$) and PPE (puff neurons: $r = 0.72$, $p < 0.05$; no-puff neurons: $r = 0.3$, $p =$ n.s.). These findings suggest that striosomes have sufficient information about RPE and PPE to provide teaching signals for downstream circuits.

We also tested whether there were differences in the proportion of sSPNs and mSPNs with activity related to action–outcome associations. By regression analysis, we did not find a difference (Supplementary Fig. 5j–l). This result suggests that SPNs in both compartments can represent associations between actions and outcomes. Thus, our findings indicate that outcome and prediction errors signaling is stronger in sSPNs, but that responses related to movement per se are similarly detected in each population.

Outcome decoding reliability is greater for sSPN activity than for mSPN activity. The single-cell SPN data indicated that sSPNs

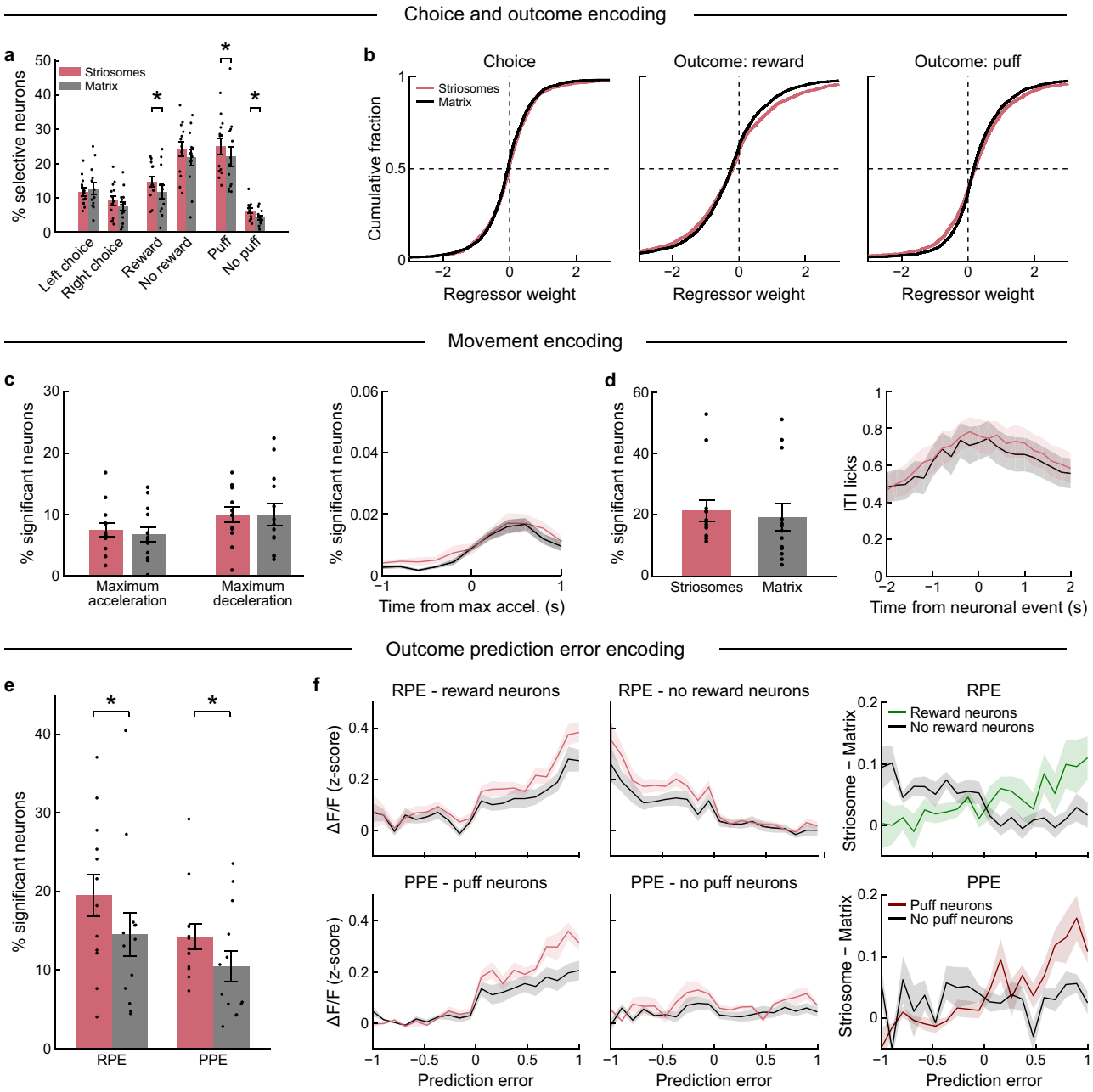


Fig. 5 Preferential outcome and PE representation for reward and punishment in striosomes. a Percentage (mean ± SEM) of neurons selective for chosen action, reward outcome or puff outcome. More sSPNs were activated by reward (sSPNs: 14.7 ± 1.4%, mSPNs: 11.7 ± 1.9%, $p = 0.039$, $t = 2.30$, df = 12, two-sided repeated measures t-test, $n = 13$ mice), puff (sSPNs: 24.9 ± 2.2%, mSPNs: 22.0 ± 2.8%, $p = 0.046$, $t = 2.23$, df = 12, $n = 13$ mice) and no-puff (sSPNs: 6.1 ± 0.7%, mSPNs: 4.2 ± 0.6%, $p = 0.017$, $t = 2.77$, df = 12, $n = 13$ mice) outcomes. There was a trend for no-reward neurons ($p = 0.07$) and no difference for action-selective neurons. $^*p < 0.05$. **b** Regression coefficients of sSPNs ($n = 2249$) and mSPNs ($n = 3582$). Their distribution was significantly different for reward ($p = 0.0044$) and puff outcomes ($p = 0.0091$) but not for chosen action (two-sided Kolmogorov–Smirnov test). **c** Neuronal activity aligned to maximum wheel acceleration and deceleration during ITIs. Percentage (mean ± SEM) of sSPNs/mSPNs with significant movement modulation (left), and average responses (mean ± SEM) of modulated neurons, aligned to peak acceleration (right; acceleration: sSPNs = 7.4 ± 1.1%, mSPNs = 6.7 ± 1.2%; $n = 13$ mice, two-sided unpaired t-test, $p = 0.67$, $t = -0.43$, df = 24; deceleration: sSPNs: 9.9 ± 1.3%, mSPNs: 10.0 ± 1.8%, two-sided unpaired t-test, $p = 0.99$, $t = 0.02$, df = 24; $n = 13$ mice). **d** Licking during ITIs aligned to detected transients. Panels show percentage (mean ± SEM) of sSPNs/mSPNs with significantly more activity-triggered licking than expected by chance (left; sSPNs: 20.4 ± 3.7%, mSPNs: 18.6 ± 4.2%; $n = 13$ mice, two-sided unpaired t-test, $p = 0.76$, $t = 0.31$, df = 24, $n = 13$ mice), and licking aligned to neuronal activity (mean ± SEM) for significantly modulated neurons (right). **e** Percentage of sSPNs/mSPNs modulated by RPE (sSPNs: 19.5 ± 2.7% (mean ± SEM), mSPNs: 14.5 ± 2.8%, $p = 0.011$, $t = 2.98$, df = 12, two-sided repeated measures t-test, $n = 13$ mice) or PPE (sSPNs: 14.2 ± 1.6, mSPNs: 10.4 ± 1.9%, $p = 0.039$, $t = 2.32$, df = 12, $n = 13$ mice) in the stepwise regression model. $^*p < 0.05$. **f** RPE/PPE modulation of $\Delta F/F$ in reward- or puff-selective sSPNs/mSPNs. Correlation coefficients were higher for sSPNs than for mSPNs (reward: $p = 0.042$; puff: $p = 0.0089$, $n = 13$ mice, two-sided repeated measures t-test). Right: the difference between the modulation in the sSPNs and mSPNs was correlated with RPE/PPE in reward ($r = 0.77$; $p < 0.01$, $n = 13$ mice, Pearson correlation) and puff ($r = 0.72$; $p < 0.05$, $n = 13$ mice) neurons, respectively. Data shown represent mean ± SEM. Source data are provided as a Source Data file.

exhibited preferential outcome-related activity, but that there was no clear compartmental difference in encoding of motor behavior. Yet it was still possible that differences between striosome and matrix compartments in movement-related activity would emerge when analyzing population activity. We therefore used decoding analyses to evaluate which task-relevant information could be read out by downstream structures from the striatal population activity. We trained artificial neural networks (ANNs) to classify trials with each of the eight possible action–outcome combinations. We trained separate models using all SPNs or only sSPNs or mSPNs (Fig. 6a). We first performed a decoding analysis using pseudo-trials, constructed by concatenating the activity of SPNs recorded across all sessions ($n = 75$ sessions). Pseudo-trial-based models predicted the action–reward–puff combination with very high accuracy (Fig. 6b).

Because pseudo-trials can inflate decoding accuracy by decoupling behavioral and neuronal variability, we also trained ANNs for individual sessions. There were 47 sessions in which all of the 8 trial combinations (chosen action × reward × puff) occurred more than 20 times. For these sessions, we randomly sampled 8 × 20 trials for decoding analysis. We could accurately decode the action–outcome combination in 49% of the trials (chance level with 8 categories: 12.5%) using the activity of all SPNs within a session ($n = 77.75 \pm 4.16$ neurons, 47 sessions; Fig. 6c), and 43% and 37%, respectively, using matching numbers of sSPNs and mSPNs ($n = 28.68 \pm 1.82$ neurons, 47 sessions; Supplementary Fig. 6a, b). We quantified the percentage of trials in which the chosen action, the reward outcome or the puff outcome was misclassified (Fig. 6c–e). As expected, mSPN-based models had significantly more misclassifications for the reward ($p < 0.001$, $t = 3.81$, $df = 46$) and puff ($p < 0.001$, $t = 6.51$, $df = 46$) outcome than sSPN-based models, but there was no difference for chosen action (Fig. 6e). The higher accuracy for sSPNs was observed regardless of the total number of transients that were observed in a session, indicating that differences in transient rates cannot account for the differential accuracy (Fig. 6f).

These decoding analyses demonstrate that SPN ensemble activity contains robust information about action–outcome contingencies, and they confirm the conclusion based on our previous analyses (Fig. 5) that the sSPNs that we imaged exhibited stronger activity in relation to the outcome of actions than did the mSPNs imaged simultaneously in the same fields of view.

Decoding future actions using SPN activity. Updating of action values allows agents to adapt their behavior in order to maximize value. We tested whether the recorded activity contained information about future decision-making. Some SPNs responded differentially depending on whether the mouse was going to stay with its action or switch in the next trial (Fig. 7a). We tested for switch/stay encoding in single neurons by chi-square analysis. More SPNs significantly encoded future switching than stay behavior during the outcome period (sSPNs: $p < 0.001$, $t = 6.21$, $df = 12$; mSPNs: $p < 0.001$, $t = 5.72$, $df = 12$; Fig. 7b). In agreement with the hypothesis that striosomes are important for cost-benefit decision-making[16,51], we found significantly more sSPNs than mSPNs encoding future switching selectively in trials with a combined reward and puff outcome (switch: $p < 0.05$, $t = 2.50$, $df = 12$; stay: $p < 0.05$, $t = 2.64$, $df = 12$).

We further assessed the representation of future behavior by regression analyses. First, we identified SPNs encoding reward-puff-switching information (Fig. 7c). Second, we performed a time-resolved analysis in which we identified factors predicting activity for every SPN for every time point (Fig. 7d). In this analysis, the percentage of neurons with activity representing future switch/stay behavior was low relative to other factors throughout the trial and was not significantly different for sSPNs and mSPNs.

Because we found only weak representations of future switching in single SPNs, we again used ANNs to test whether ensemble SPN activity could predict future switching. Some trial types were extremely rare (e.g., no reward–puff–stay). We included 51 sessions that had at least 5 trials for all of the 8 reward–puff–switch combinations. Despite the limited samples used for training (32 training trials and 8 test trials), we could decode reward and puff outcomes and predict future switch/stay behavior with a higher accuracy than expected by chance, which for 8 categories corresponds to 12.5% (Fig. 7e, f, Supplementary Fig. 7a, b). The sSPN model again had lower misclassification rates than the mSPN model for decoding reward ($p < 0.001$, $t = 4.08$, $df = 12$) and puff ($p < 0.001$, $t = 6.00$, $df = 12$) outcome.

Finally, we asked whether SPN activity in the 3 s preceding trial start was predictive of the next action. Models trained with the pseudo-trial activity of all SPNs predicted future actions with high accuracy. ANN models trained using single sessions had a lower accuracy but one still greater than chance (50%; Supplementary Fig. 7c). Together, these population-decoding analyses demonstrate that the striatal ensemble activity during outcomes and before trial start contained information about the future behavior of the animal.

Discussion

Our findings in mice demonstrate that single projection neurons in the anterodorsal striatum can represent in their activity associations between an action and both rewarding and aversive outcomes of that action. We observed this action–outcome encoding by implementing a new bandit task in which mice learned without explicit cuing or instruction to maximize reward and simultaneously minimize punishment. In this task, the block sizes deliberately consisted of relatively few trials so that the animals continuously adapted their behavior in response to evolving prediction errors. Within this CBB task context, visually identified SPNs in striosomes exhibited enhanced encoding of RPE and PPE, relative to nearby matrix SPNs imaged in the same field of view. These findings emphasize multiplexed encoding of action–outcome representations in the striatum and support a differential function of striosomes in underpinning behavioral adaptation in environments requiring assessment of cost and benefit.

We analyzed the Ca^{++} responses of identified SPNs within an RL computational framework. In conventional RL models[1], the consequence of an action is usually evaluated by a scalar. Following the concept of RL theories, it is thus tempting to assign the positive value for good outcomes and the negative value for bad outcomes by integrating cost and benefit for the evaluative outcome signal. However, when the animals receive multiple modalities for the outcome (i.e., reward and airpuff), neuronal activity can represent the outcome value as a single scalar or can represent the outcomes individually for each modality, as we observed in our sample of SPNs.

The CBB task developed here was an attempt to resolve this ambiguity, as was done previously in humans in a task requiring subjects to maximize monetary rewards and minimize electrical shocks[56]. Our findings suggest that in the SPNs imaged, outcomes were not represented as a single scalar value. Rather, associations between actions and qualitatively different outcomes were represented in a multiplexed manner in partially overlapping populations of SPNs. We divided SPNs for our analysis into 'value' and 'non-value' neurons, and found approximately equal numbers of the two. In this task, the outcomes differed both in

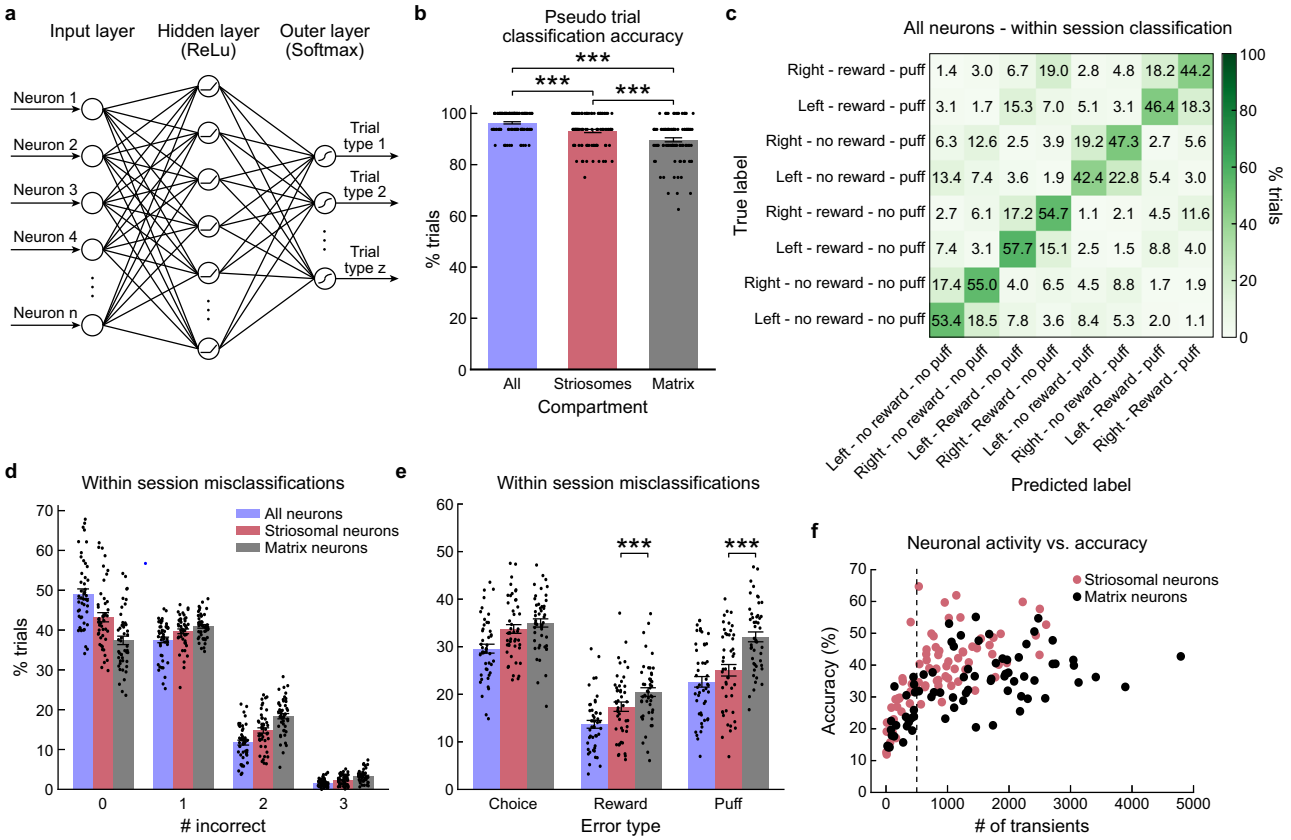


Fig. 6 Decoding analysis of action–outcome combinations. a Artificial neural network architecture used for behavioral decoding. The hidden layer had the same size as the input layer. **b** Accuracy (mean ± SEM) of model using pseudo-trials for all SPNs ($n = 5831$) and subsampled sSPNs ($n = 2249$) or mSPNs ($n = 2249$). The model including all SPNs had a higher accuracy than the compartment-specific models (all vs. matrix: $t = 7.30$, $p = 1e{-}12$; all vs. striosomes: $t = 4.18$, $p = 0.000043$, $n = 100$ repetitions, two-sided repeated measures t-test, $^{***}p < 0.001$), and the striosome model outperformed the matrix model ($t = 3.4$, $p = 0.0008$) (all SPNs: 96%; sSPNs: 93%; mSPNs: 90%). **c** Confusion matrix of session-based models showing percentage of trials with the true and predicted label for each of the 8 different trial types. **d** Percentage of trials in which 0, 1, 2 or 3 dimensions were incorrectly predicted by session-based models (mean ± SEM, $n = 47$). **e** The models based on sSPN activity significantly outperformed the models based on mSPN activity when predicting reward (misclassifications: sSPNs: 17.4 ± 1.0% (mean ± SEM); mSPNs: 20.5 ± 0.9%, two-sided repeated measures t-test, $p = 0.00041$, $t = 3.81$, df = 46, $n = 47$) and puff (misclassifications: sSPNs: 25.1 ± 1.2%; mSPNs: 32.1 ± 1.0%, two-sided repeated measures t-test, $p = 1e{-}8$, $t = 6.51$, df = 46, $n = 47$) outcome but not chosen action (sSPNs: 33.7 ± 0.9% and mSPNs: 35.0 ± 0.9%, $n = 47$). The model based on all neurons had an accuracy higher than the matrix model for all three features (choice: $t = 6.46$, $p = 0.000065$; reward: $t = 5.45$, $p = 1e{-}7$; puff: $t = 6.16$, $p = 1e{-}8$; $p < 0.001$, two-sided repeated measures t-test, $n = 47$) and higher than the striosome model for chosen action ($t = 3.12$, $p = 0.002$, two-sided repeated measures t-test, $n = 47$) and reward ($t = 2.86$, $p = 0.0053$, two-sided repeated measures t-test, $n = 47$), but the performance was not significantly different for puff ($^{***}p < 0.001$). **f** Relationship between the model accuracy and the total number of transients recorded across all imaged neurons per session. Some sessions had a lower number of transients, due to low activity or a low number of neurons. Sessions on the left of the dashed line were excluded from the decoding analysis. Source data are provided as a Source Data file.

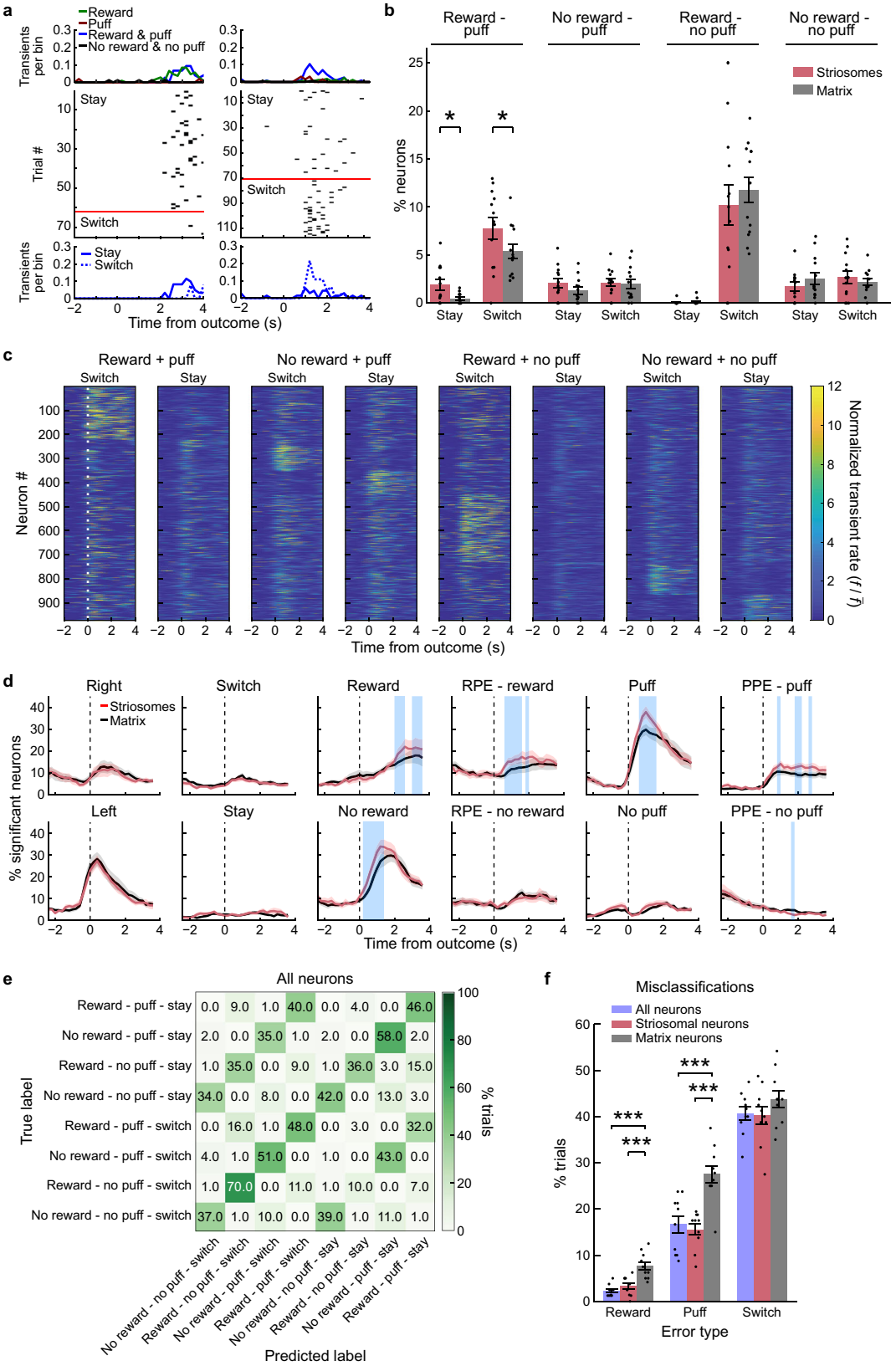
terms of their valence and identity; it remains to be tested whether different outcomes with the same valence are also represented in parallel.

It is possible that the neurons responding to rewards and puffs, or to the absence of both, could signal the overall salience of the outcome. However, we consider this unlikely. Salience of an outcome is not selective for an action, whereas outcome activity often was. Salience also is not specific to the direction of a prediction error. Very large negative or positive RPE/PPE could be equally salient, but we did not observe SPNs that were active for both unexpected reward delivery and unexpected reward omission (or puff delivery/omissions).

For both reward and puff outcome-sensitive neurons, we noted that there were more neurons responding to negative outcomes (reward omission and puff delivery). This bias is perhaps related to the specific imaging location, or to negative events being more behaviorally relevant, among potential biases; our data could not be used to directly support either hypothesis.

Because of the extended signaling inherent to Ca^{++} imaging, we also were unable to resolve whether the SPN activity during the decision period reflected the expected outcome or the total expected value. Nor were we able to analyze subthreshold inhibitory responses: whereas transients can be reliably detected and are known to be tightly linked to action potential firing, other types of calcium dynamics, including reductions in intracellular Ca^{++}, could not reliably be resolved. However, the 2-photon imaging preparation that we employed had the great advantage of allowing us to access at a single-cell level the activities of SPNs, and to visually identify their striosomal and matrix identity for over 5000 simultaneously imaged sSPNs and mSPNs.

Pioneering work has shown that direct and indirect pathway neurons (dSPNs and iSPNs) modulate, respectively, approach and avoidance behavior[20,57]. Lacking intersectional genetics, we were not able to examine this critical differential representation of outcomes by identified dSPNs and iSPNs. What we could do, however, was to provide a template of information about the



attributes of SPNs clustered together in visually identified striosome and matrix compartments.

Visual identification of these SPN subtypes is critical, because genetic models that label striosomes have labeling in the matrix (false discovery rate: ~25%), label striosomes sparsely (false negative rate: ~70%) and are highly biased to striosomal neurons expressing D1 receptors[49,58,59]. Our model also suffers from this problem[52–54], but has the advantage of labeling the neuropil in the birthdate-labeled striosomes, which made it possible to identify visually the striosomal modules with high reliability[53].

We also emphasize that the results here were all obtained by imaging fields in the anterodorsal striatum, including the region

Fig. 7 Neuronal representations of future switch/stay behavior. a Two examples of neurons showing activity in relation to upcoming switch/stay behavior in reward/puff trials. Top: average activity in trials with different reward–puff outcome combinations. Middle: raster plots of all reward–puff trials followed by the same action or the opposite action (above or below the red line, respectively). Bottom: average activity of the neurons in reward–puff trials followed by staying (solid) or switching (dashed). **b** Percentage (mean ± SEM) of neurons that significantly differentiated future staying or switching in trials with different outcome combinations ($n = 13$ mice). In reward–puff trials, significantly more sSPNs were active in relation to future stay and switch behavior (switch: $p = 0.028$, $t = 2.50$, df = 12, $n = 13$; stay: $p = 0.022$, $t = 2.64$, df = 12, $n = 13$, two-sided repeated measures t-test). $^*p < 0.05$. **c** Activity of all neurons showing switch/stay selectivity, averaged per trial type (reward × puff × switch/stay). **d** Time-resolved stepwise regression model showing the average percentage of neurons per mouse that had individual factors included in that timepoint-specific model. Blue shading indicates time points with a significant difference in the percentage of sSPNs and mSPNs ($p < 0.05$, two-sided paired t-test, $n = 13$ mice). Data shown represent mean ± SEM. **e** Confusion matrix for decoding of reward-puff-switch/stay trials in a pseudo-trial analysis. **f** The sSPN and combined models outperformed the mSPN model when decoding reward (misclassifications: sSPNs: 3.4 ± 0.7% (mean ± SEM), mSPNs: 7.8 ± 0.9%, $t = 5.04$, $p = 0.00070$, two-sided repeated measures t-test, $n = 10$) or puff (sSPNs: 15.6 ± 1.2%, mSPNs: 27.6 ± 1.6%, $t = 5.54$, $p = 0.00036$, two-sided repeated measures t-test, $n = 10$) outcome but not switch/stay behavior. $^{***}p < 0.001$. There were no differences between the sSPN and combined models. Source data are provided as a Source Data file.

with a prominent necklace-like crescent of striosomes (AP range: −0.2 to 1.2 mm, ML range: 1 to 2.4 mm); within this region, we did not observe noticeable differences between the way actions, rewards and puff outcomes activated the SPNs, with the caveat that in most mice the range in which we imaged was smaller and we could not directly compare lateral and medial regions, thought to have different encoding properties[7,10,60–62]. Recordings were all made from the left hemisphere. We did not detect differences in the activity related to ipsilateral/contralateral actions (Fig. 2d, Supplementary Figs. 2e–h, 5a, b); but we do not rule out the possibility of undetected differences in action–outcome encoding between the two hemispheres.

sSPNs are considered to be the main source of striatal input to dopamine-containing neurons of the substantia nigra (SN), with mSPNs mainly projecting to the non-dopaminergic neurons of the SN[45–47,49,63]. This parallel innervation recalls the parallel network structure of the actor-critic architecture in RL models[4,40–44,64,65]. We found that it was in relation to reinforcement-related factors that the striosomal and matrix SPNs differed: sSPNs had more pronounced responses to the reward and puff outcomes of given actions, particularly when RPE and PPE were high.

Our finding that sSPNs are biased to encode RPE and PPE expands on the previous reports of the activity of sSPNs and mSPNs in Pavlovian tasks, wherein striosomal activity dominated during reward cues but not in the outcome period[53,66]. In Pavlovian tasks, the cues, but not the outcomes, are relevant for updating value expectations. In bandit tasks, outcomes are critical for updating action values. In this task implementing costs and benefits as reinforcers, striosomes and matrix were often co-active, but striosomal activity dominated when new information was provided that resulted in behavioral adaptations, either because of rewarding or aversive consequences. Our findings are thus important conceptually in distinguishing potentially different experimental conditions calling up the activity of striosomes and matrix, as well as in understanding the physiology of potential 'critic' circuits.

As previously reported[67], we also observed tdTomato-positive neurons in the matrix. It was not possible to compare this population with the general striosomal population with sufficient statistical power due to low numbers (32 neurons). This low percentage partly reflects our sampling strategy, in which we chose fields-of-view with clearly delineated striosome and matrix compartments, and the chosen timing of Tamoxifen injection to result in sparse, but highly reliable, striosomal labeling.

We did not observe stronger responses related to motor behavior in mSPNs than in sSPNs, as often hypothesized. This result at first was surprising, but could be due to the selection of imaged regions relative to innervation by motor cortex and related cortical regions[49,68]. This result is, however, aligned with the view that the striatum is important for learning the value of actions. We did not observe many neurons that encoded only detected motor behavior without regard to the outcome.

Our findings point to synergistic, cooperative patterns of activation of striosomal and matrix pathways. Such synergism was originally proposed for the D1-expressing direct and D2-expressing indirect pathways[69], and more recent evidence strongly aligns itself with this view of the direct–indirect pathway control system[70,71]. We suggest that not only the direct–indirect axis of striatal organization, but also the striosome–matrix axis, likely exhibit both synergistic and opposing activity patterns depending on the environmental contingencies requiring adaptive behavioral response.

We draw two central conclusions from this work. First, many outcome-related SPN activities in the anterodorsal striatum could not be readily accounted for by a value-coding scheme. Instead, the activities are capable of representing multiple independent action–outcome contingencies, indicating a multiplex coding framework. Second, striosomes and their nearby surrounding matrix exhibit marked differences in their representations of outcomes, with striosomes biased toward responding in relation to information that is critical for learning processes. These findings provide a window into how the striatum and its compartmental divisions contribute to adaptive behaviors guided by rewards and punishments in uncertain environments.

Methods

Experimental model and subjects. All experimental procedures performed on mice were approved by the Massachusetts Institute of Technology Committee on Animal Care. We used 13 Mash1-CreER (het) × Ai14 (het) mice for recording striatal activity, which were offspring of female Mash1-CreER (het) × Ai14 (homo) mice crossed with male C57Bl/6J mice. The female transgenic mice were from a colony that were generated by crossing Mash1(Ascl1)-CreER mice[72] (Ascl1tm1.1(Cre/ERT2)Jejo/J, Jackson Laboratory) with Ai14-tdTomato Cre-dependent mice[73] (B6;129S6-Gt(ROSA)26Sor, Jackson Laboratory), which were then crossed with FVB mice to improve breeding results. To generate mice with striosome labeling, we timed the breeding. We paired one male C57Bl/6J mouse with 2 female Mash1-CreER × Ai14 mice. Labeling was induced by injecting pregnant dams with Tamoxifen, dissolved in corn oil, by oral gavage (100 mg/kg) at embryonic day 11.5[53,54]. By this method, Mash1 is expressed predominantly in future striosomal neurons.

We studied 6–10-week-old mice of both sexes (9 females and 4 males). Mice were housed singly after the first surgery.

Virus injections and surgery. We prepared mice for behavioral training and 2-photon Ca^{++} imaging using previously described procedures[53]. Virus was injected in the striatum of adult mice to express GCaMP6s using aseptic stereotaxic surgery, with mice deeply anesthetized by 3% isoflurane and mounted in a stereotaxic frame. Anesthesia was maintained with constant 1–2% isoflurane, adjusted as needed. Meloxicam (1 mg/kg) and slow-release buprenorphine (1 mg/kg) were given subcutaneously to provide analgesia. The head was shaved and cleaned with depilatory cream (Nair), and surgical areas were disinfected with three alternating applications of povidone-iodine and 70% ethanol. The skin was incised to expose the skull, and the head was leveled to align bregma and lambda in the z-axis. Two

burr holes were drilled in the skull, and 500 nl of AAV5-hSyn-GCaMP6s-WPRE-SV40 virus was injected through each opening made at the following coordinates defined relative to bregma: 1) 0.1 mm anterior, 1.9 mm lateral, 2.7 mm ventral; and 2) 0.9 mm anterior, 1.7 mm lateral, and 2.5 mm ventral. Injections were made over 10 min, and the needle was left in place for another 10 min. The incision was sutured shut, and mice were kept warm while they recovered from the surgery. Wet food and meloxicam were provided for 3 days to facilitate post-surgical recovery.

Mice were surgically implanted with an imaging cannula at 10–14 days after virus injection. The imaging cannula was assembled by adhering a 2.7 mm glass coverslip to the end of a custom stainless-steel metal tube that was adapted to provide extra stability (1.7 mm long, 2.7 mm diameter eMachineshop; 3D design available at request). Mice were water-restricted for a week before the cannula surgery, which significantly improved the clarity of the preparation[74]. Mice were deeply anesthetized with isoflurane and head-fixed in the stereotaxic setup. Bregma and lambda were aligned in the *z*-axis, and craniotomy coordinates were marked at 0.6 mm anterior and 2.1 mm lateral to bregma. The skull was tilted and rolled by 5° to make it more horizontal at the cannula implant site. A trephine dental drill was used to make a 2.7 mm diameter craniotomy. The cortical tissue was aspirated with gentle suction under constant perfusion with sterile 0.01 M phosphate buffered saline until the underlying white matter appeared. A small layer of the white matter was removed and covered with a thin layer of Kwik-Sil (World Precision Instruments) before inserting the cannula into the cavity. The cannula, as well as a custom head plate, was attached to the skull using Metabond (Parkell). Pre- and post-surgical analgesia regimen and care were as described above for the virus surgery.

Behavioral apparatus and task training. The behavioral rig was constructed from optical hardware (Thorlabs) and custom 3D-printed parts (designs available upon request). Mice were head-fixed with their forepaws resting on a Lego wheel, which they could rotate freely, and they reported their decisions by rotating the wheel left or right. The wheel was coupled to a rotary encoder, allowing us to register its rotations with high temporal resolution. The rotary encoder was connected to a microcontroller (Arduino), which ran a custom routine that sampled the position of the rotary encoder every 10 ms. In the event of a movement, the microcontroller sent a timestamp and the amount of movement to a behavioral control computer. The behavioral task was implemented with custom software written in MATLAB (MathWorks) using the Data Acquisition toolbox. Water rewards (3–7 µl, dependent on the behavior of the individual mice) were delivered through a tube by opening a solenoid value for a calibrated period. Air puffs (20 psi) were delivered through a tube positioned on the snout and were similarly controlled with a solenoid valve. Both solenoid valves were located outside the imaging setup. Licking was measure via a conductance-based method[75].

We trained mice through successive stages of shaping in order to implement the final task. Mice were water-restricted for at least one week after recovering from the surgery and received ~1 ml of water per day. If mice did not earn their water allotment during the task, they were given hydrogel (Clear H_2O) in their home cage. Mice were first habituated to head-fixation and trained to lick the water tube to receive water rewards. When mice licked reliably, training started. Trial start was signaled with an auditory tone (4 kHz), after which mice had 3 s to move the wheel by 15° in either direction for it to count as a response. An ITI of 3.5 s was used between trials. To initiate a trial, mice had to hold the wheel still for 2 s. If wheel movement exceeded 5° in either direction during this time, then an additional 1 s was added to the delay. In the first stage of training, movements to either direction resulted in water reward delivery and no air puff. Reward was delivered when a complete response was registered and signaled with an auditory tone (10 kHz). When mice made responses in at least 150 trials on consecutive days, they moved on to the next training stage, in which only one action (turning the wheel to the right or left direction) was paired with a reward, with contingencies changing after 6–15 rewarded trials. Mice progressed to the next stage of training when they made complete responses in at least 80% of trials, performed ~200 trials, and showed minimal signs of side bias. In the next stage, air puffs were introduced. In every block, one action was associated with a puff with 100% probability. When mice avoided 6–15 puffs by choosing the action that was not paired with the air puff, the action–air puff contingency switched. Mice progressed to the final task after reaching the same criteria as above.

In the final task, each action was probabilistically (80%) associated with a water reward and an air puff outcome in blocks of trials. For example, in a right puff block, left and right actions were punished with a 0% and 80% probability, respectively. Reward block transition occurred after 6–15 rewards were delivered (chosen randomly for each block). Similarly, puff blocks transitioned after mice avoided 6–15 puffs by selecting the action not associated with puff. Hence, reward and puff block transitions occurred independently so that in some trials the same action was associated with 80% reward and puff probability and in some trials the opposite actions. At the beginning of each session (i.e., the first reward and puff blocks), reward and puff outcomes were independently assigned to one action. At block transitions, the sides associated with reward or puff were switched.

Once mice made responses in at least 150 trials, had no bias and showed clear switch/stay behavior dependent on reward and puff outcomes, they were moved to a behavioral rig under a 2-photon microscope. In most cases, moving from the previous behavioral setup required training the mice for additional 2–7 days until they reached the performance criterion again. The training duration was between 6 weeks and 3 months, with a clear dependence on initial starting age.

Imaging. Two-photon Ca^{++} imaging procedures were as described previously[53]. Briefly, GCaMP6s and tdTomato fluorescence was imaged through a LUMPlan FL, ×40, 0.8NA objective using galvo-galvo scanning with a Prairie Ultima IV 2-photon microscopy system (Bruker). Excitation light at 910 nm was provided by a tunable Ti:Sapphire laser equipped with dispersion compensation (Mai Tai Deep See, Spectra-Physics). Green and red fluorescence emission signals were split with a dichroic mirror (Semrock) and directed to GaAsP photomultiplier tubes (Hamamatsu). Images were acquired at a frame rate of 5 Hz. Laser power at the sample ranged from 11 to 42 mW, depending on the imaging depth and level of GCaMP6s expression. We selected fields of view (FOVs) that allowed simultaneous imaging of striosomal and matrix neurons. The FOVs had both clearly labeled GCaMP6s-expressing cells in striosomes, as defined by dense tdTomato signal in the neuropil, as well as in areas free of tdTomato labeling. Cells were classified as striosomal or matrix depending on whether they were found inside the tdTomato-expressing neuropil zones. In total, we imaged 75 FOVs in 13 mice.

Image processing

Realignment. We motion-corrected imaging videos by realigning all images to an average reference frame. We first realigned all images in the red stationary channel to the average of all frames in that channel using 2-dimensional cross-correlation (template matching and slice alignment plugin)[76]. Next, we realigned the images in the green channel on the basis of the frame-by-frame translations that were calculated for the red channel. We previously found that this procedure does not differentially affect the registration of striosomes and matrix[53].

Detection of regions of interest and extraction of ΔF/F. After registration, we detected neurons manually based on the mean, standard deviation and maximum projections. Custom MATLAB scripts were used to calculate local background fluorescence surrounding the somatic regions of interest. The background fluorescence, multiplied by 0.7, was subtracted from the somatic signals, as described previously[77]. After this, the baseline fluorescence for all neurons (F_0) was calculated using *K*-means (KS)-density clustering to estimate the mode of the fluorescence distribution. $\Delta F/F$ was calculated as $\Delta F/F = (F_t - F_0)/F_0$.

Detection of striosomes. Striosomes were visually identified in the plane imaged as regions with dense labeling of tdTomato in the neuropil. We did not record in regions with tdTomato-positive neurons but no clear labeling in the striosomal neuropil. Every cell was classified as striosomal or matrix on the basis of its location in the visually identified compartments. tdTomato-positive neurons outside of the striosomes were included in the striosomal population (32 out of 296 tdTomato-labeled neurons), based on observation that there are neurons in the matrix[67] that have some characteristics in common with those of striosomes. Including these neurons or not did not affect our results. In total, 5831 neurons were recorded, of which 2249 neurons were classified as striosomal.

Detection of transients. We used a custom algorithm for detecting Ca^{++} transient onsets in the *Z*-scored $\Delta F/F$. Transients were scored if they met several criteria. First, the size had to be at least 5 times the standard deviation higher than the median $\Delta F/F$. The derivative of the signal also had to exceed the standard deviation, which resulted in detecting onsets. In addition, the mean $\Delta F/F$ signal in the 1-s period following the onset, subtracted by the signal in a 0.6-s window before the onset, had to be bigger than 2 times the standard deviation. After this, in cases where multiple sequential time samples had detected transients, we only kept the first as the event onset. This simple algorithm efficiently detected events (Supplementary Fig. 2a).

Behavioral modeling

Cost–benefit RL model. We adapted existing RL models[1,78] to include both rewarding and aversive outcomes. We formulated two models. In the first, *Q*-values for reward and punishment are learned in parallel, and during the decision both are integrated. In the second, there is one set of *Q*-values, based on the weighed value of an outcome.

Parallel RL model: The inferred expectations about reward and puff outcomes linked to actions were modeled using two sets of *Q*-values, one set for reward and one for puffs. In every trial, the decision was modeled using a sigmoid function based on the relative *Q*-values for reward and the relative *Q*-values for puff. The sensitivity of the mouse to these differences was parameterized by β_{rew} and β_{puff}. A bias term β_0 was also included. After observing the outcomes, the *Q*-values for both reward and puff were updated in parallel, using the same rules as in existing models based on reward only. RPE and PPE were calculated by comparing the outcome with the expected value of the chosen action. *Q*-values were updated for both the chosen and the unchosen action. A total of 2 × 3 different learning rates was applied (chosen action followed by outcome: $\alpha_{rew/puff}$; chosen action, no reward/puff delivered $\alpha_{unrew/nopuff}$; non-chosen action: $\gamma_{rew/puff}$). Bayesian information criterion analysis showed that including all six learning rates improved the model.

One model was created for every mouse. The model was fit using the fminsearch function in MATLAB. Performance was evaluated using 5-fold cross validation.

Integrated RL model: This model is similar to the two set model, but only one set of *Q*-values is learned. After every trial, the reward and puff outcomes are combined into one outcome value.

The prediction error is then calculated using this value. Next, the single set of *Q*-values is updated and used to make a decision in the next trial.

Auto-regressive behavioral model. We used an autoregressive model to quantify how strongly rewards, puff and actions impact future decision-making (Eq. (8); Fig. 1g). We fit a model in which we predicted actions in trial i with the five previous trials (lagging by j). There were eight sets of predictors. For both left and right (subscript l/r), there were N (no reward and no puff), R (reward, no puff), P (puff, no reward) and B (both reward and puff). For every trial, one of these was scored as 1 if that particular combination occurred, and the other 7 were scored as 0. A bias term β_0 was also included. The model was fit for every mouse using the glmfit function in MATLAB. Performance was evaluated using 5-fold cross validation.

$$\begin{aligned} Y(i) = &\sum_{j=1}^{n} \beta_j^{\text{No reward/puff}} (N_r(i-j) - N_l(i-j)) \\ &+ \sum_{j=1}^{n} \beta_j^{\text{Reward}} (R_r(i-j) - R_l(i-j)) \\ &+ \sum_{j=1}^{n} \beta_j^{\text{Puff}} (P_r(i-j) - P_l(i-j)) \\ &+ \sum_{j=1}^{n} \beta_j^{\text{Both}} (B_r(i-j) - B_l(i-j)) + \beta_0 \end{aligned} \quad (8)$$

$$P^{\text{R}}(i) = \frac{1}{1 + e^{-Y(i)}}$$

$$P^{\text{L}}(i) = 1 - P^{\text{R}}(i)$$

Analysis of single-neuron activity

Analysis of action, reward and puff encoding. To identify neurons that were selective for chosen action, reward or puff outcomes, we used two approaches. First, we used chi-square analysis to test for selectivity by comparing the number of trials with at least one Ca^{++} transient across two conditions (e.g., left vs. right action). To find neurons that were active in relation to multiple features, we performed sequential chi-square tests. First, we tested the individual factors (left vs. right, reward vs. no reward, puff vs. no puff). In neurons that had activity selective for a particular trial type, we then took only these trials, split them again for another factor and ran another chi-square test. Sequential statistical testing is conservative as neurons have to repeatedly pass a statistical criterion. We therefore used a regression as second approach to identify neurons representing actions, outcomes and combinations of these.

We also used a stepwise logistic regression analysis to define neuron types. In these analyses, we selected a number of predictors and then searched for the optimal model for every neuron to predict whether it has a transient in a given trial or not. In every iteration, the effect of adding or removing every possible factor from the model was evaluated. The change that resulted in the largest explanatory power of the model was selected if adding/removing a factor significantly improved the performance of the model. The null hypothesis for the statistical test was that the coefficient of that factor had a coefficient of zero if it would be included in the model. The p-value that was used for cutoff was 0.05. For this analysis, we used the stepwiseglm function in MATLAB.

Partial regression analysis was performed to confirm the effect of adding prediction errors to the models that explain neuronal activity on the basis of trial outcomes. For reward, puff and combined outcomes, we first performed regression of the neurons' activity against, for example, reward. We then regressed RPE against reward, and then finally we regressed the residuals from the first regression against the residuals from the second regression. Then the number of neurons in which this last regression was significant was determined to quantify the ability of RPE, PPE, and combined PE to predict neuronal activity.

We performed logistic regression analysis for every neuron to quantify the strength of the relationship between chosen action, reward outcome and puff outcome with the probability of having a transient in a trial. To avoid overfitting, we used L2 regularization.

In all cases, we first calculated the mean per mice and performed statistics on these data. Therefore $n = 13$ mice for all analyses except stated otherwise.

RPE and PPE representations. We analyzed whether neuronal activity in reward/no-reward or puff/no-puff neurons was modulated by RPE and PPE. We used the RL model to infer prediction errors for every trial and then binned these. We then calculated for the different neuronal populations the amplitude of the transients in Z-scored $\Delta F/F$. To test whether neuronal activity was modulated by the prediction errors, we performed correlation analyses. This was done for all trials and also only including the trials with the outcome that the neurons encoded. Statistics were done on the mean activities per mouse ($n = 13$).

Comparing direction selectivity for action–reward and action–puff associations. To test the hypothesis that neurons encode outcome with opposite valence for different actions, we compared the direction selectivity for rewarding and aversive outcomes. First, we used a chi-square test to find neurons that were selective for specific action × reward and action × puff combinations. We then determined the joint distribution of action × reward and action × puff selective neurons.

Selectivity analysis. We compared the selectivity of reward and puff activity in the neurons that were selectively active for both reward and puff outcomes. For this comparison, we calculated for every neuron the proportion of reward/no-reward/puff/no-puff trials in which transients occurred. The selectivity index for reward was calculated by dividing this proportion for trials with a reward by the sum of the proportions of trials with a reward and trials without a reward, resulting in values between 0 and 1, where 1 means that all transients occur in trials with a reward and 0 means that all transients occur in trials without a reward. We did the same for puff selectivity. We transformed these data by subtracting 0.5 and multiplying the outcome by 2 to have a range of −1 to 1, with 0 meaning that transients are as likely to occur in trials with or without the outcome.

Wheel movement and licking analysis. We tested how neurons represented wheel movements and licking during the task and in ITIs. Wheel movement and licking data were first binned at 5 Hz to facilitate comparison with the GCaMP6s neuronal signals. For detecting wheel movement bouts, we first smoothed the absolute value of the whole-session movement trace using a 7-point moving average. This facilitated detecting events preceded by ~1 s of no movement. The resulting trace was baseline adjusted using the ksdensity function in MATLAB and binarized using a wheel movement threshold of 0.44°. Event onset and offset times were determined as timepoints at which the binarized movement trace shifted from 0 to 1 and from 1 to 0, respectively. The smoothing introduced a lag of 2 time bins, which was corrected in the onset/offset times. This analysis detects all movement bouts occurring during the session. To restrict analysis to movements occurring during the ITI, we removed all bouts with onset or offset occurring during a 3-s period from trial start, as well as bouts that had a trial occurring in the middle of them. Individual bouts were labeled as leftward or rightward based on the mean direction of movement. Neuronal data were aligned to time of peak acceleration or deceleration of the wheel within movement bouts. For peak acceleration analysis, significance was determined by comparing the number of movement bouts with neuronal transients occurring in a 1-s window before or after the maximum acceleration using a chi-square test. Significance testing for deceleration was done similarly, except that activity occurring between 0.4 s before and 0.6 s after peak deceleration was compared to a preceding 1-s baseline period.

We performed two types of licking analysis. First, we determined how often licking was coincident with neuronal activity during the ITI period. For each neuron, we generated a peri-stimulus time histogram over a 4-s window by aligning licks to the time of detected neuronal transients. Licks occurring during 3 s after trial start were not counted. Only neurons with at least 10 transients during the session were included in this analysis. We used a shuffle test to generate a null distribution of licking expected by chance. We realigned licking on random permutations of neuronal transient times. This process was repeated 100 times. The total number of licks computed from the observed data was compared to the resulting null distribution. Neurons with observed values outside of the center 95% of the null distribution were considered significant. In a second analysis, we detected licking bouts using a procedure similar to that described for the wheel movements above, with a few differences. The whole-session licking trace was smoothed with a 5-point moving average, and binarization was performed with a lick threshold of 0.2. No baseline adjustment was necessary. For whole-session analysis, neuronal activity was aligned to times of lick bout onsets. For the ITI analysis, only licking bouts occurring outside of trials were used, as described above. Significance testing for both analyses was done by comparing the 1-s pre-licking and 1-s post-licking periods using a chi-square test. For both wheel movement and licking analysis, data from all neurons for individual mice were concatenated together, and the reported values were based on number of mice.

Decoding analysis. We used artificial neural networks to decode behavior from the neuronal activity. This approach resulted in higher accuracy than support vector machines or logistic regression, and it allowed us to decode multiple classes in one analysis without having to create multiple one-versus-rest models. The models consisted of three layers. The input layer used the activity of every neuron. The hidden layer was the same size as the input layer and used ReLU activation functions. The output layer had one node per target and used a softmax function to calculate a probability distribution. We used the Adam optimizer and a learning rate of 0.001, and trained the network in 100 epochs. Dropout (0.3) was used to prevent overfitting. Parameters were chosen on the basis of a grid search.

We created models that were based on pseudo-trials and models in which only activity from single sessions was included. We created pseudo-trials by selecting from every neuron a fixed number of trials for each target and concatenated across neurons. The number of trials was set separately for the different analysis based on how often trials typically occurred. In the decoding of trial state (action and both outcomes), we used 20 trials resulting in 47 sessions in which all of the 8 trial combinations (chosen action × reward × puff) had more than 20 trials. The other

sessions, in which certain combinations did not occur often enough, were excluded. For reward × puff × switch/stay decoding, we used 5 trials per trial type, which was the minimum for 5-fold cross validation and resulted in 51 sessions. For decoding left/right actions based on ITI activity, we included 85 trials, so that all sessions could be included. The pseudo-trial analysis was performed 100 times, every time taking different trials from the sessions.

Pseudo-trials have the advantage that one can combine neurons from different sessions, which increases the predictive power of the model. However, the number of trials has to be restricted to ensure enough trial samples from all, or most, neurons. In addition, pseudo-trials decouple behavioral and neuronal variability. Therefore, we also created models for each session using the same parameters as described above. To compare striosomal and matrix populations, we always took the same number of neurons from every session. The largest population was therefore subsampled. This procedure was repeated 40 times for each analysis.

The models were implemented using TensorFlow in Python 3. Other libraries used were NumPy, SciPy, and Pandas.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Source data for all figures are provided with this paper. Data have not been deposited in an online repository since we did not find a data repository with a standard that is widely adopted by the research community for these data types. Requests for additional data can be directed to the corresponding author and will be attempted to be handled within two weeks. Source data are provided with this paper.

Code availability

Code for fitting the cost–benefit reinforcement learning model has been deposited on GitHub (https://github.com/bloemb/CBC_RL_model). Custom code used in this study is available from the corresponding author upon reasonable request.

Received: 17 August 2021; Accepted: 15 February 2022;
Published online: 22 March 2022

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 2018).
2. Adams, C. D. & Dickinson, A. Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. Sect. B* **33B**, 109–121 (1981).
3. Gremel, C. M. & Costa, R. M. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nat. Commun.* **4**, 2264 (2013).
4. O'Doherty, J. et al. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
5. Simon, N. W., Wood, J. & Moghaddam, B. Action-outcome relationships are represented differently by medial prefrontal and orbitofrontal cortex neurons during action execution. *J. Neurophysiol.* **114**, 3374–3385 (2015).
6. Smith, K. S. & Graybiel, A. M. Habit formation. *Dialogues Clin. Neurosci.* **18**, 33–43 (2016).
7. 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).
8. Yang, L. & Masmanidis, S. C. Differential encoding of action selection by orbitofrontal and striatal population dynamics. *J. Neurophysiol.* **124**, 634–644 (2020).
9. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *Eur. J. Neurosci.* **22**, 505–512 (2005).
10. Yin, H. H. et al. Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nat. Neurosci.* **12**, 333–341 (2009).
11. Lau, B. & Glimcher, P. W. Action and outcome encoding in the primate caudate nucleus. *J. Neurosci.* **27**, 14502–14514 (2007).
12. Stalnaker, T. A., Calhoon, G. G., Ogawa, M., Roesch, M. R. & Schoenbaum, G. Neural correlates of stimulus-response and response-outcome associations in dorsolateral versus dorsomedial striatum. *Front. Integr. Neurosci.* **4**, 12 (2010).
13. Walters, C. J., Jubran, J., Sheehan, A., Erickson, M. T. & Redish, A. D. Avoid-approach conflict behaviors differentially affected by anxiolytics: implications for a computational model of risky decision-making. *Psychopharmacology* **236**, 2513–2525 (2019).
14. Amemori, K. I. & Graybiel, A. M. Localized microstimulation of primate pregenual cingulate cortex induces negative decision-making. *Nat. Neurosci.* **15**, 776–785 (2012).
15. Aupperle, R. L. & Paulus, M. P. Neural systems underlying approach and avoidance in anxiety disorders. *Dialogues Clin. Neurosci.* **12**, 517–531 (2010).
16. Friedman, A. et al. A corticostriatal path targeting striosomes controls decision-making under conflict. *Cell* **161**, 1320–1333 (2015).
17. Wallis, J. D. & Rushworth, M. F. S. Integrating benefits and costs in decision making. In *Neuroeconomics* (ed. Glimcher, P.W. & Fehr, E.) Ch. 22 (Academic Press, 2014).
18. Delgado, M. R., Li, J., Schiller, D. & Phelps, E. A. The role of the striatum in aversive learning and aversive prediction errors. *Philos. Trans. R. Soc. B: Biol. Sci.* **363**, 3787–3800 (2008).
19. Hikida, T., Kimura, K., Wada, N., Funabiki, K. & Nakanishi, S. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron* **66**, 896–907 (2010).
20. Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–818 (2012).
21. Palminteri, S. et al. Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* **76**, 998–1009 (2012).
22. Stephenson-Jones, M. et al. Opposing contributions of GABAergic and glutamatergic ventral pallidal neurons to motivational behaviors. *Neuron* **105**, 921–933 (2020).
23. Averbeck, B. B. & Costa, V. D. Motivational neural circuits underlying reinforcement learning. *Nat. Neurosci.* **20**, 505–512 (2017).
24. Hamid, A. A. et al. Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2015).
25. Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N. & Komiyama, T. Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858–1872 (2019).
26. Lau, B. & Glimcher, P. W. Value representations in the primate striatum during matching behavior. *Neuron* **58**, 451–463 (2008).
27. Neftci, E. O. & Averbeck, B. B. Reinforcement learning in artificial and biological systems. *Nat. Mach. Intell.* **1**, 133–149 (2019).
28. Nonomura, S. et al. Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. *Neuron* **99**, 1302–1314 (2018).
29. Parker, N. F. et al. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nat. Neurosci.* **19**, 845–854 (2016).
30. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340 (2005).
31. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782–1787 (2004).
32. 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**, 1281–1289 (2012).
33. Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L. & Nyberg, L. Transfer of learning after updating training mediated by the striatum. *Science* **320**, 1510–1512 (2008).
34. Matamales, M. et al. Local D2- to D1-neuron transmodulation updates goal-directed learning in the striatum. *Science* **367**, 549–555 (2020).
35. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
36. Menegas, W., Akiti, K., Amo, R., Uchida, N. & Watabe-Uchida, M. Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nat. Neurosci.* **21**, 1421–1430 (2018).
37. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
38. Schultz, W. Dopamine reward prediction-error signalling: a two-component response. *Nat. Rev. Neurosci.* **17**, 183–195 (2016).
39. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
40. Brown, J., Bullock, D. & Grossberg, S. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *J. Neurosci.* **19**, 10502–10511 (1999).
41. Doya, K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Curr. Opin. Neurobiol.* **10**, 732–739 (2000).
42. Houk, J. C., Adams, J. L. & Barto, A. G. *A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement* (The MIT Press, 1995).
43. Joel, D., Niv, Y. & Ruppin, E. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Netw.* **15**, 535–547 (2002).
44. Takahashi, Y., Schoenbaum, G. & Niv, Y. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model. *Front. Neurosci.* **2**, 86–99 (2009).
45. Crittenden, J. R. et al. Striosome-dendron bouquets highlight a unique striatonigral circuit targeting dopamine-containing neurons. *Proc. Natl Acad. Sci. USA* **113**, 11318–11323 (2016).

46. Evans, R. C. et al. Functional dissection of basal ganglia inhibitory input onto SNc dopaminergic neurons. *Cell Rep.* **32**, 108156 (2020).
47. Fujiyama, F. et al. Exclusive and common targets of neostriatofugal projections of rat striosome neurons: a single neuron-tracing study using a viral vector. *Eur. J. Neurosci.* **33**, 668–677 (2011).
48. Matsushima, A. & Graybiel, A. M. Combinatorial developmental controls on striatonigral circuits. *Cell Rep.* **31**, 107778 (2020).
49. McGregor, M. M. et al. Functionally distinct connectivity of developmentally targeted striosome neurons. *Cell Rep.* **29**, 1419–1428 (2019).
50. Amemori, S. et al. Microstimulation of primate neocortex targeting striosomes induces negative decision-making. *Eur. J. Neurosci.* **51**, 731–741 (2020).
51. Friedman, A. et al. Chronic stress alters striosome-circuit dynamics, leading to aberrant decision-making. *Cell* **171**, 1191–1205 (2017).
52. Friedman, A. et al. Striosomes mediate value-based learning vulnerable in age and a huntington's disease model. *Cell* **183**, 918–934 (2020).
53. Bloem, B., Huda, R., Sur, M. & Graybiel, A. M. Two-photon imaging in mice shows striosomes and matrix have overlapping but differential reinforcement-related responses. *Elife* **6**, e32353 (2017).
54. Kelly, S. M. et al. Radial glial lineage progression and differential intermediate progenitor amplification underlie striatal compartments and circuit organization. *Neuron* **99**, 345–361 (2018).
55. Amemori, K. I., Gibb, L. G. & Graybiel, A. M. Shifting responsibly: the importance of striatal modularity to reinforcement learning in uncertain environments. *Front. Hum. Neurosci.* **5**, 47 (2011).
56. Seymour, B., Daw, N. D., Roiser, J. P., Dayan, P. & Dolan, R. Serotonin selectively modulates reward value in human decision-making. *J. Neurosci.* **32**, 5833–5842 (2012).
57. LeBlanc, K. H. et al. Striatopallidal neurons control avoidance behavior in exploratory tasks. *Mol. Psychiatry* **25**, 491–505 (2020).
58. Banghart, M. R., Neufeld, S. Q., Wong, N. C. & Sabatini, B. L. Enkephalin disinhibits mu opioid receptor-rich striatal patches via delta opioid receptors. *Neuron* **88**, 1227–1239 (2015).
59. Miyamoto, Y., Katayama, S., Shigematsu, N., Nishi, A. & Fukuda, T. Striosome-based map of the mouse striatum that is conformable to both cortical afferent topography and uneven distributions of dopamine D1 and D2 receptor-expressing cells. *Brain Struct. Funct.* **223**, 4275–4291 (2018).
60. Lee, J., Wang, W. & Sabatini, B. L. Anatomically segregated basal ganglia pathways allow parallel behavioral modulation. *Nat. Neurosci.* **23**, 1388–1398 (2020).
61. Lerner, T. N. et al. Intact-brain analyses reveal distinct information carried by SNc dopamine subcircuits. *Cell* **162**, 635–647 (2015).
62. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189 (2004).
63. Gerfen, C. R. The neostriatal mosaic: compartmentalization of corticostriatal input and striatonigral output systems. *Nature* **311**, 461–464 (1984).
64. Stephenson-Jones, M., Kardamakis, A. A., Robertson, B. & Grillner, S. Independent circuits in the basal ganglia for the evaluation and selection of actions. *Proc. Natl Acad. Sci. USA* **110**, 3670–3679 (2013).
65. Suri, R. E. TD models of reward predictive responses in dopamine neurons. *Neural Netw.* **15**, 523–533 (2002).
66. Yoshizawa, T., Ito, M. & Doya, K. Reward-predictive neural activities in striatal striosome compartments. *eNeuro* **5**, 1–14 (2018).
67. Smith, J. B. et al. Genetic-based dissection unveils the inputs and outputs of striatal patch and matrix compartments. *Neuron* **91**, 1069–1084 (2016).
68. Kincaid, A. E. & Wilson, C. J. Corticostriatal innervation of the patch and matrix in the rat neostriatum. *J. Comp. Neurol.* **374**, 578–592 (1996).
69. Mink, J. W. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* **50**, 381–425 (1996).
70. Tecuapetla, F., Matias, S., Dugue, G. P., Mainen, Z. F. & Costa, R. M. Balanced activity in basal ganglia projection pathways is critical for contraversive movements. *Nat. Commun.* **5**, 4315 (2014).
71. Tecuapetla, F., Jin, X., Lima, S. Q. & Costa, R. M. Complementary contributions of striatal projection pathways to action initiation and execution. *Cell* **166**, 703–715 (2016).
72. Kim, E. J., Ables, J. L., Dickel, L. K., Eisch, A. J. & Johnson, J. E. Ascl1 (Mash1) defines cells with long-term neurogenic potential in subgranular and subventricular zones in adult mouse brain. *PLoS ONE* **6**, e18472 (2011).
73. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2009).
74. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016).
75. Slotnick, B. A simple 2-transistor touch or lick detector circuit. *J. Exp. Anal. Behav.* **91**, 253–255 (2009).
76. Tseng, Q. et al. A new micropatterning method of soft substrates reveals that different tumorigenic signals can promote or reduce cell contraction levels. *Lab Chip* **11**, 2231–2240 (2011).
77. Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
78. 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).

Acknowledgements

We thank Dr. Tomoko Yoshida, Erik D. Nelson and Dr. Christian Wuethrich for help with histology, and Dr. Yasuo Kubota for help in preparing the figures and manuscript. This work was supported by National Institute of Mental Health (R01 MH060379 to A.M.G.; R00 MH112855 to R.H.), Saks Kavanaugh Foundation (to A.M.G.), William N. & Bernice E. Bumpus Foundation (RRDA Pilot: 2013.1 to A.M.G.; Postdoctoral Fellowship to B.B.), Simons Foundation (306140 to A.M.G.), Nancy Lurie Marks Family Foundation (to A.M.G.), National Eye Institute (R01 EY028219 and R01 EY007023 to M.S.), National Institute of Neurological Disease and Stroke (U01 NS090473 to M.S.), National Science Foundation (EF1451125 to M.S.), Simons Foundation Autism Research Initiative (to M.S.), Army Research Office (W911NF-21-1-0328 to M.S. and A.M.G.) and JSPS KAKENHI (20H03555, 20H05469, 20H05063, and 18K19497 to K.A.).

Author contributions

B.B., R.H., K.A., and A.M.G. conceptualized the study; B.B. and R.H. performed formal analysis; A.M.G. performed informal analysis; B.B., R.H., A.A., G.K., A.L.W., and C.W.C. performed experiments; B.B., R.H., and A.M.G. developed methodology; B.B., R.H., K.A., and A.M.G. wrote the manuscript; B.B., R.H., K.A., M.S., and A.M.G. reviewed and edited the manuscript; B.B., R.H., M.S., and A.M.G. acquired funding; M.S. and A.M.G. supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28983-5.

Correspondence and requests for materials should be addressed to Ann M. Graybiel.

Peer review information *Nature Communications* thanks Wolfgang Kelsch, Hugo Tejeda and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Inventory of Supporting Information

Title: Multiplexed action-outcome representation by striatal striosome-matrix compartments detected with a mouse cost-benefit foraging task

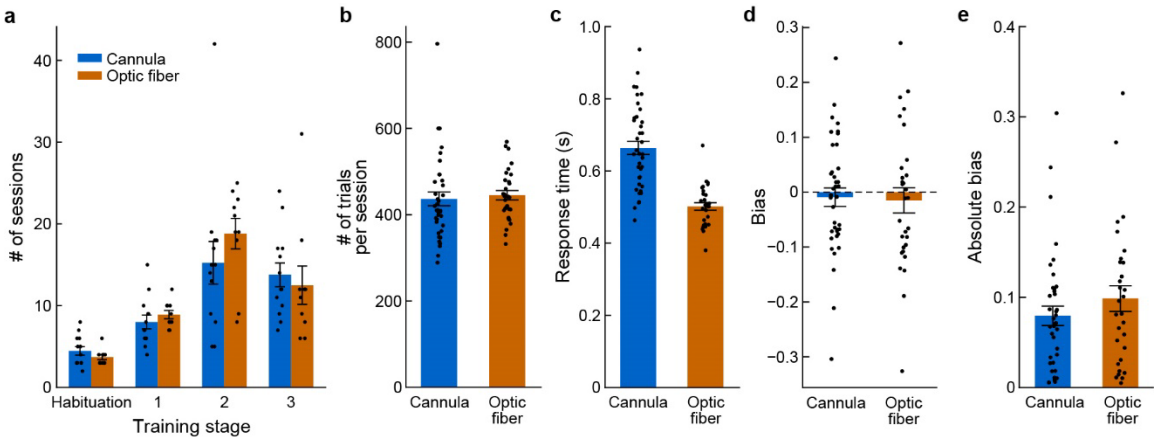
Authors: Bernard Bloem, Rafiq Huda, Ken-ichi Amemori, Alexander Abate, Gaya Krishna, Anna Wilson, Cody W. Carter, Mriganka Sur, Ann M. Graybiel

Supplemental Figures:

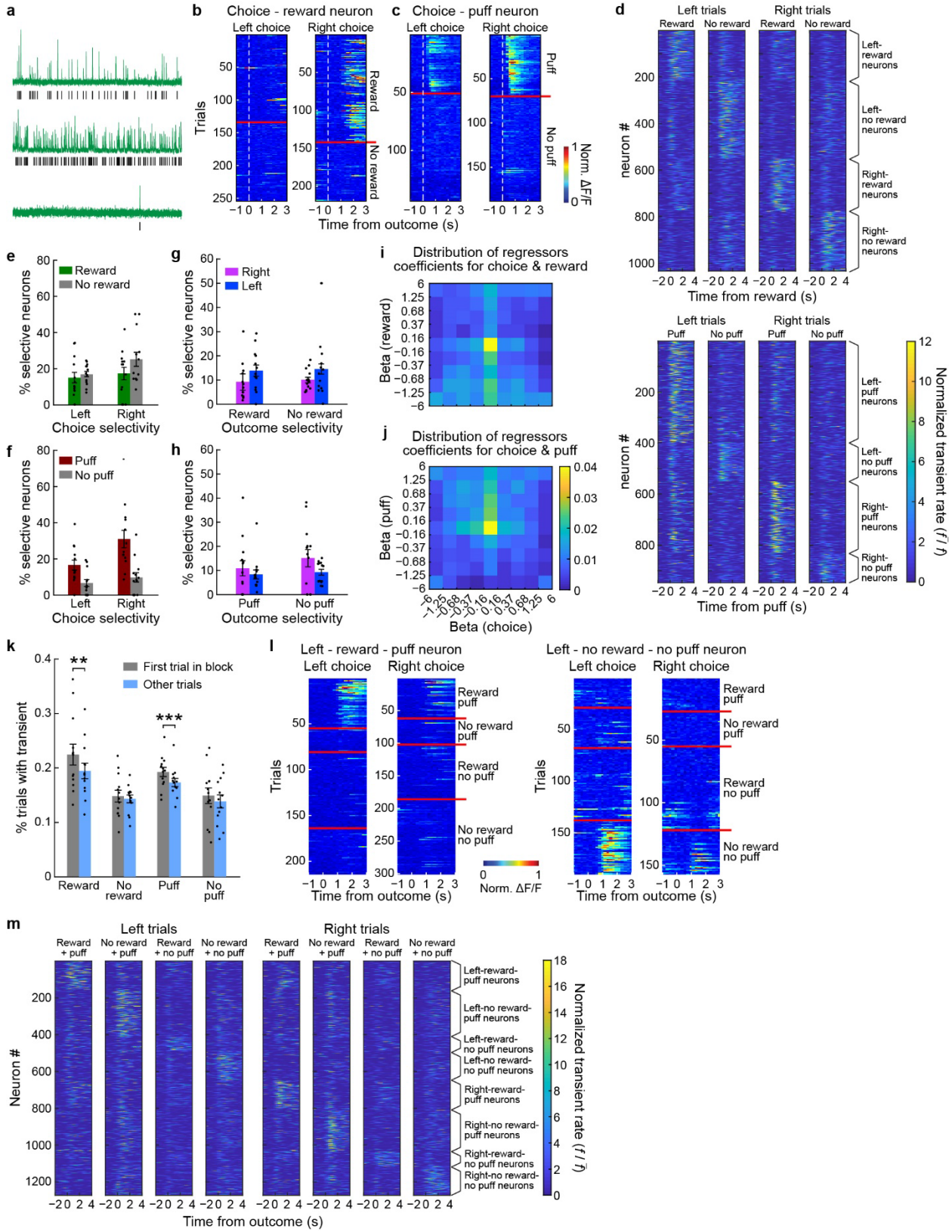
Additional supplemental information:

Source data: source_data.zip
Cost-benefit reinforcement learning model: https://github.com/bloemb/CBC_RL_model

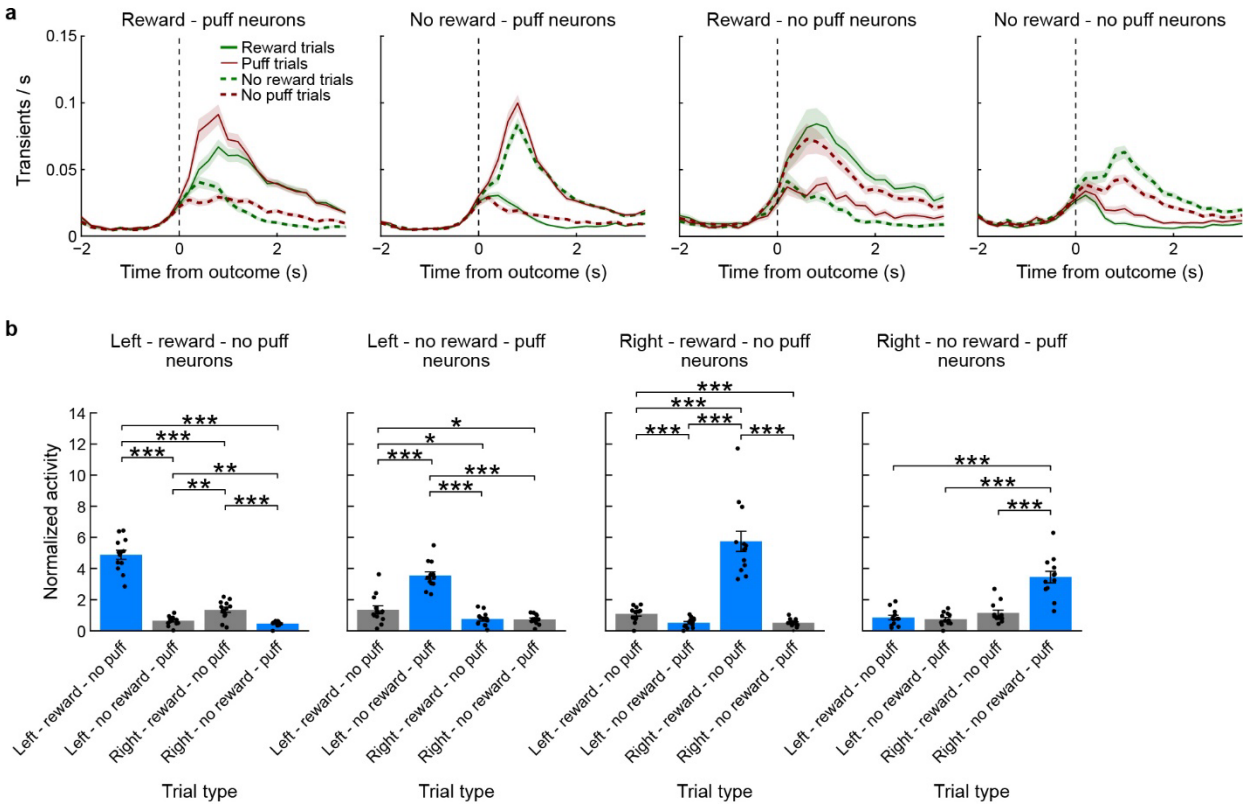
SUPPLEMENTARY FIGURES



Supplementary Fig. 1 Effect of cannula implantation on behavioral performance. a Number of sessions that was required to progress through all training stages, as described in Methods (mean ± SEM, n = 10 control mice and 12 mice with cannula). **b** Average number of trials performed in the last three sessions before reaching the final performance criterion (mean ± SEM, control: n = 30, cannula: n = 36). **c** Average response time in the last three sessions before reaching criterion (mean ± SEM, control: n = 30, cannula: n = 36). **d** Average bias in the last three sessions before reaching criterion (mean ± SEM, control: n = 30, cannula: n = 36). **e** Average absolute bias in the last three sessions before reaching criterion (mean ± SEM, control: n = 30, cannula: n = 36). Source data are provided as a Source Data file.

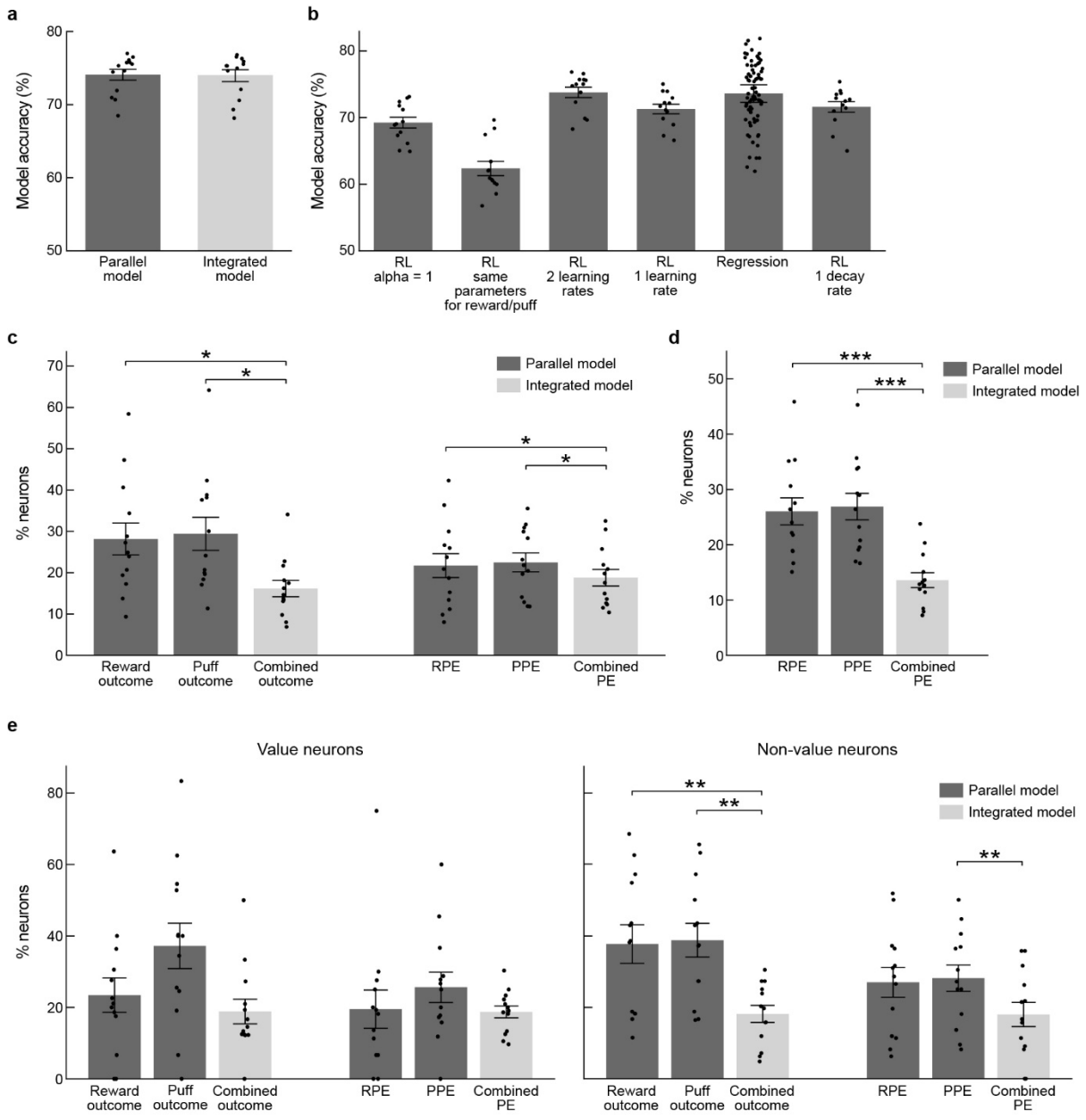

a
b
Choice - reward neuron
Left choice
Right choice
Trials
Reward
No reward
Time from outcome (s)
c
Choice - puff neuron
Puff
No puff
Norm. ΔF/F
d
Left trials
Right trials
Reward
No reward
neuron #
Left-reward neurons
Left-no reward neurons
Right-reward neurons
Right-no reward neurons
Time from reward (s)
Puff
No puff
Left-puff neurons
Left-no puff neurons
Right-puff neurons
Right-no puff neurons
Normalized transient rate (f / f̄)
Time from puff (s)
e
% selective neurons
Choice selectivity
f
g
Outcome selectivity
h
i
Distribution of regressors coefficients for choice & reward
Beta (reward)
j
Distribution of regressors coefficients for choice & puff
Beta (puff)
Beta (choice)
k
% trials with transient
First trial in block
Other trials
Reward
No reward
Puff
No puff
l
Left - reward - puff neuron
Left - no reward - no puff neuron
Reward puff
No reward puff
Reward no puff
No reward no puff
m
Left trials
Right trials
Reward + puff
No reward + puff
Reward + no puff
No reward + no puff
Neuron #
Left-reward-puff neurons
Left-no reward-puff neurons
Left-reward-no puff neurons
Left-no reward-no puff neurons
Right-reward-puff neurons
Right-no reward-puff neurons
Right-reward-no puff neurons
Right-no reward-no puff neurons
Time from outcome (s)

Supplementary Fig. 2 Action-outcome association representations by SPNs. a ΔF/F fluorescent traces of three sample neurons recorded over 30 min (green) and the time of detected Ca^{++} events (black). **b, c** Two examples of neurons showing activity selective for action-reward (**b**) and action-puff (**c**) combinations. Trials are shown (rows) separately for left/right action, with red lines demarcating reward/no-reward or puff/no-puff outcome trials. **d** The analysis for Fig. 2e was repeated, but with half of the trials used for detecting the neuronal response types and the other half for calculating the average response in the different trial types. The results confirm the validity of the detected neuron types. **e** Percentage (mean ± SEM) of action-selective neurons with selectivity for reward or no-reward outcomes (left: reward = 14.9 ± 2.9%, no reward = 16.8 ± 1.6%; right: reward = 17.2 ± 3.4%, no reward = 25.0 ± 3.9%; n = 13 mice). There were no significant main effects or interactions (ANOVA). **f** Percentage of action-selective neurons with selectivity for puff or no-puff trials (left: puff = 16.6 ± 2.7%, no puff = 6.6 ± 1.9%; right: puff = 30.9 ± 4.8%, no puff = 9.7 ± 2.6%; mean ± SEM, n = 13 mice). There were significant main effects of puff outcome ($p = 0.000024$) and choice ($p = 0.012$; ANOVA). **g** Percentage of reward-outcome-selective neurons with selectivity for left or right actions. No significant main or interaction effects were detected (mean ± SEM , n = 13 mice). **h** Percentage of puff-outcome-selective neurons that was selective for the two actions. No significant effects were detected (mean ± SEM, n = 13 mice). **I, j** Joint distribution of chosen action and reward (**i**) or puff (**j**) regressor coefficients. Horizontal and vertical bins were chosen to divide the non-zero coefficients equally among the bins. **k** Comparison of action-outcome related responses in action-outcome selective neurons in the first trial after a block switch versus other trials with the same action and outcome (mean ± SEM, n = 13 mice, ***$p < 0.001$). **l** Two examples of neurons showing activity representing an association between an action and both reward and puff outcomes. **m** The analysis of Fig. 2h repeated with half of the trials used for detecting the neuron types and the other half for calculating the average responses. Source data are provided as a Source Data file.



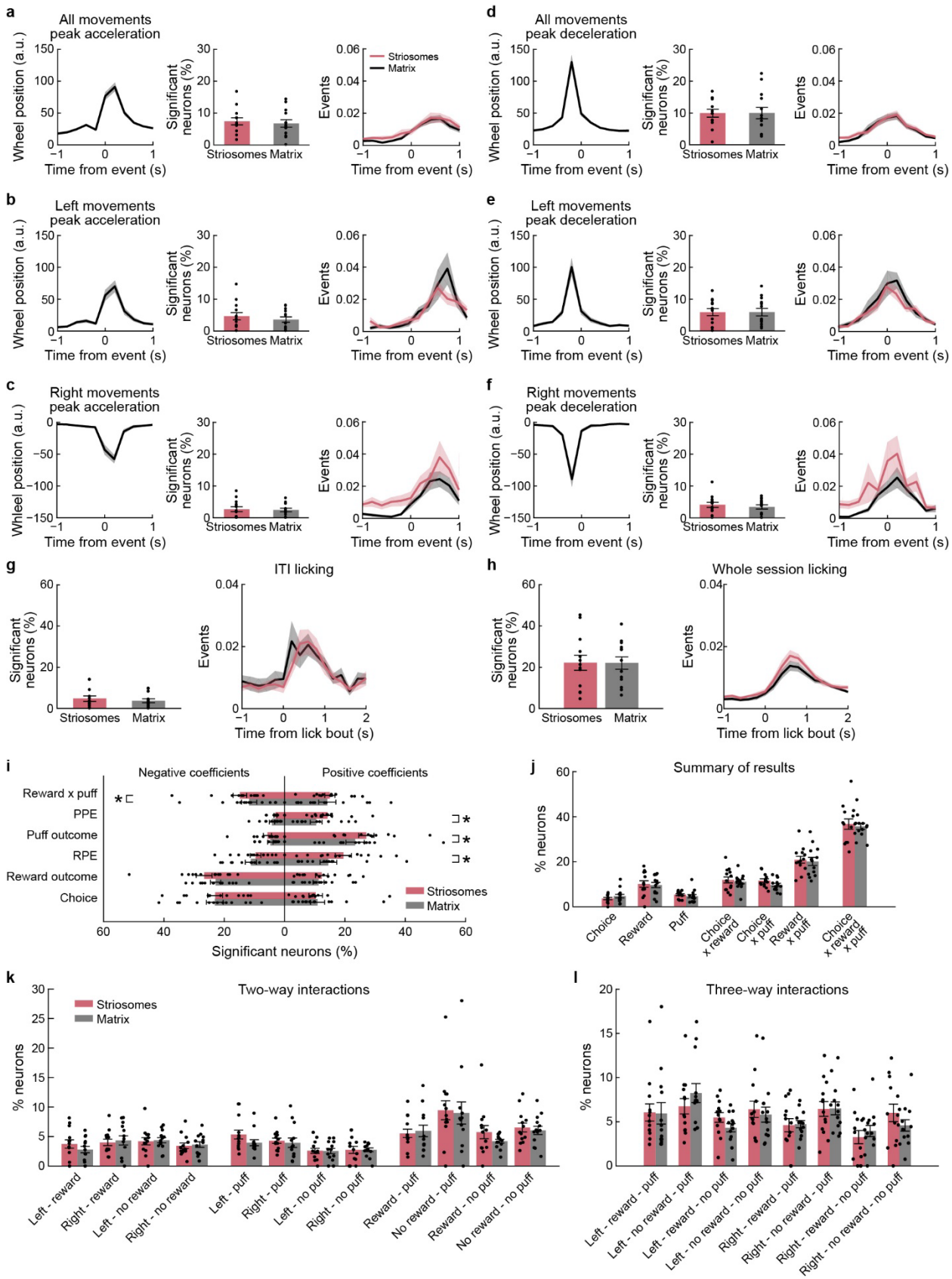
Supplementary Fig. 3 Activity and selectivity of reward/no-reward/puff/no-puff neurons. a Activity (mean ± SEM) of four groups of neurons with activity selective for different outcome combinations in reward/no-reward/puff/no-puff trials (n = 13 mice). **b** Normalized activity of neurons with value-like responses across different trial types (mean ± SEM, n = 13 mice). Neurons that were active in trials in which a good outcome was delivered for one action did not have enhanced activity when the other action was paired with a bad outcome, or vice versa. For all 4 types of neurons, ANOVA indicated significant main effects and interactions ($p < 0.001$). Two-sided post-hoc t-test showed significance between different trial types for each neuron group (Left - no reward - puff neurons: left - no reward - puff trials - left - reward - no puff trials $p = 0.00050$; Left - no reward - puff neurons: left - no reward - puff trials - right - no reward - puff trials $p = 0.025$; Left - no reward - puff neurons: left - no reward - puff trials - right - reward - no puff trials $p = 0.041$; Left - no reward - puff neurons: left - reward - no puff trials - right - no reward - puff trials $p = 1e{-}06$; Left - no reward - puff neurons: left - reward - no puff trials - right - reward - no puff trials $p = 1e{-}07$; Left - reward - no puff neurons: left - no reward - puff trials - left -

reward - no puff trials p = 1e−09; Left - reward - no puff neurons: left - no reward - puff trials - right - no reward - puff trials p = 1e−07; Left - reward - no puff neurons: left - no reward - puff trials - right - reward - no puff trials p = 1e−09; Left - reward - no puff neurons: left - reward - no puff trials - right - no reward - puff trials p = 0.0064; Left - reward - no puff neurons: left - reward - no puff trials - right - reward - no puff trials p = 0.0078; Left - reward - no puff neurons: right - no reward - puff trials - right - reward - no puff trials p = 0.00036; Right - no reward - puff neurons: left - no reward - puff trials - right - reward - no puff trials p = 0.00023; Right - no reward - puff neurons: left - reward - no puff trials - right - reward - no puff trials p = 0.000055; Right - no reward - puff neurons: right - no reward - puff trials - right - reward - no puff trials p = 0.00067; Right - reward - no puff neurons: left - no reward - puff trials - left - reward - no puff trials p = 0.00075; Right - reward - no puff neurons: left - no reward - puff trials - right - no reward - puff trials p = 0.000018; Right - reward - no puff neurons: left - no reward - puff trials - right - reward - no puff trials p = 0.00045; Right - reward - no puff neurons: left - reward - no puff trials - right - no reward - puff trials p = 1e−05; Right - reward - no puff neurons: right - no reward - puff trials - right - reward - no puff trials p = 1e−05, *p < 0.05; **p < 0.01; ***p < 0.001). Source data are provided as a Source Data file.



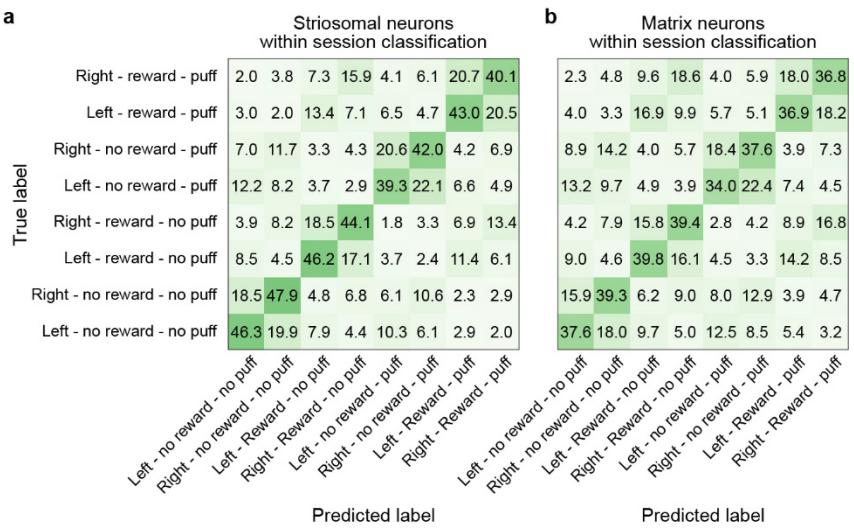
Supplementary Fig. 4 Comparison of cost-benefit reinforcement learning models. a Model accuracy of the parallel and integrated cost-benefit RL models during cross validation (mean ± SEM, n = 13 mice). **b** Cross validation accuracy of alternative simpler models. 'RL alpha = 1': a win-stay/lose-switch model was created by setting all learning rates to 1. 'RL same parameters for reward/puff': one set of learning parameters for both outcomes. 'RL 1 decay rate': there was a single decay rate for both reward and puff. 'RL 2 learning rates: the forgetting rate was set to be the same as the unlearning rate (Ito & Doya, 2009). 'RL 1 learning rate: only 1 learning rate was

used for each outcome. 'Regression': performance of a 5 trial back auto-regressive model. Data shown are mean ± SEM (n = 13 mice for the RL models and n = 75 sessions for the regression model). **c** A stepwise regression was conducted for each neuron to test which factors best account for the recorded activity. The percentage of neurons that include the factors from the two competing cost-benefit RL models was higher for the parallel model (reward outcome vs. combined outcome: $p = 0.011$; puff outcome vs. combined outcome: $p = 0.022$; RPE vs. combined PE: $p = 0.048$; PPE vs. combined PE: $p = 0.035$; $^{*}p < 0.05$, average and SEM of 13 mice; two-sided paired t-test). **d** Partial regression analysis was performed to confirm the results shown in **c** and to quantify the effect of adding prediction errors to models that explain neuronal activity on the basis of outcome. For reward, puff and combined outcomes, we first performed regression of the neurons' activity against, for example, reward. We then regressed RPE against reward, and then finally we regressed the residuals from the first regression against the residuals from the second regression (mean ± SEM, n = 13, $^{***}p < 0.001$). **e** The analysis in **c** was repeated with only neurons that responded to combinations of reward and puff outcomes, and split into 'value' and 'non-value' neurons depending on whether they responded oppositely to reward and puff outcomes or not ($^{**}p < 0.001$). Source data are provided as a Source Data file.

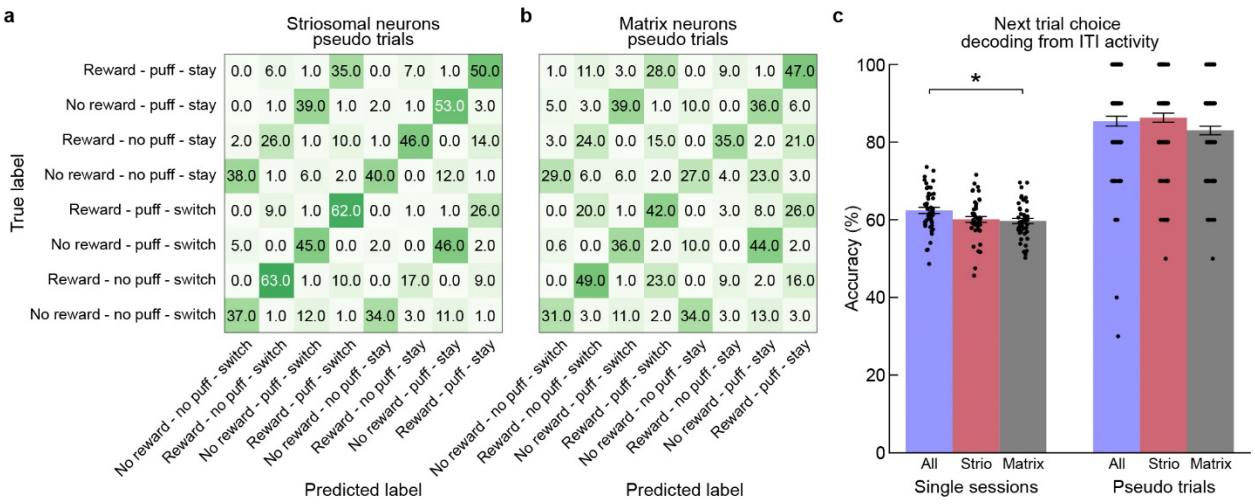

a
All movements
peak acceleration
Wheel position (a.u.)
Time from event (s)
Significant neurons (%)
Striosomes Matrix
Events
Striosomes
Matrix
d
All movements
peak deceleration
b
Left movements
peak acceleration
e
Left movements
peak deceleration
c
Right movements
peak acceleration
f
Right movements
peak deceleration
g
ITI licking
Time from lick bout (s)
h
Whole session licking
i
Negative coefficients
Positive coefficients
Reward x puff
PPE
Puff outcome
RPE
Reward outcome
Choice
Striosomes
Matrix
Significant neurons (%)
j
Summary of results
% neurons
Choice
Reward
Puff
Choice x reward
Choice x puff
Reward x puff
Choice x reward x puff
k
Two-way interactions
Striosomes
Matrix
% neurons
Left - reward
Right - reward
Left - no reward
Right - no reward
Left - puff
Right - puff
Left - no puff
Right - no puff
Reward - puff
No reward - puff
Reward - no puff
No reward - no puff
l
Three-way interactions
% neurons
Left - reward - puff
Left - no reward - puff
Left - reward - no puff
Left - no reward - no puff
Right - reward - puff
Right - no reward - puff
Right - reward - no puff
Right - no reward - no puff

Supplementary Fig. 5 Movement-related activity in sSPNs and mSPNs. a sSPN (red) and mSPN (black/gray) activity was aligned to peak acceleration of the absolute value of wheel movement bouts. Panels show average (± SEM) movements across mice (left), mean proportion (± SEM) of neurons with significant increase in movement-related activity (middle), and mean (± SEM) activity of significantly modulated neurons (right: sSPNs = 7.4 ± 1.1%, mSPNs = 6.7 ± 1.2%; n = 13 mice, two-sided unpaired t-test, p = 0.67, t = −0.43, df = 24). **b, c** Similar to **a**, except movements were divided into left (**b,** sSPNs = 4.6 ± 1.1%, mSPNs = 3.6 ± 0.8%; n = 13 mice, two-sided unpaired t-test, p = 0.47, t = −0.73, df = 24) and right (**c,** sSPNs = 2.7 ± 0.8%, mSPNs = 2.5 ± 0.6%; n = 13 mice, two-sided unpaired t-test, p = 0.81, t = −0.24, df = 24). Data are shown as mean ± SEM. **d** Similar to **a**, except activity was aligned to peak deceleration within wheel movement bouts (sSPNs: 9.9 ± 1.3%, mSPNs: 10.0 ± 1.8%; n = 13 mice, two-sided unpaired t-test, p = 0.99, t = 0.02, df = 24). Data are shown as mean ± SEM. **e, f** Similar to **b** and **c**, except with activity aligned to peak decelaration for left (**e**, sSPNs = 5.9 ± 1.1%, mSPNS = 5.9 ± 1.2%; n = 13 mice, two-sided unpaired t-test, p = 0.99, t = 0.01, df = 24) and right (**f**, sSPNs = 4.1 ± 0.8%, mSPNs = 3.4 ± 0.7%; n = 13 mice, two-sided unpaired t-test, p = 0.53, t = −0.63, df = 24) movements. Data are shown as mean ± SEM. **g** Neuronal activity (mean ±SEM) aligned to licking bout onset during ITI (sSPNs: 4.7 ± 0.9%, mSPNs: 3.7 ± 1.3; n = 13 mice, two-sided unpaired t-test, p = 0.53, t = −0.63, df = 24). **h** Same as **g**, except for licking bouts occurring during the whole session (sSPNs: 22.1 ± 3.6%, mSPNs: 22.0 ± 3.0%; n = 13 mice, two-sided unpaired t-test, p = 0.98, t = −0.02, df = 24). Data are shown as mean ± SEM. **i** Percentage (mean ± SEM) of sSPNs (red) and mSPNs (black/gray) per mouse that included the chosen action, reward and puff outcomes, their interaction, and RPE and PPE in the optimal model using stepwise regression. Significantly more sSPNs included RPE (p = 0.011, t = 2.98, df = 12), puff outcome (p = 0.049, t = 2.19, df = 12), PPE (p = 0.039, t = 2.32, df = 12) and reward x interaction (p = 0.030, t = 2.45, df = 12) in their optimal model (two-sided repeated measures t-test, n = 13, *p < 0.05, n = 13 mice). **j** Summary of stepwise regression showing average percentage of sSPNs and mSPNs per mouse with single action and outcome factors included in their optimal model, as well as different two-way and three-way interactions (mean ± SEM, n = 13). **k, l** Percentage of sSPNs and mSPNs with various two-way (**k**) and three-way (**l**) interactions included in their optimal regression model. There are no significant

differences in any of the comparisons (mean ± SEM, $p > 0.05$, $n = 13$ mice, two-sided repeated measures t-test). Source data are provided as a Source Data file.



Supplementary Fig. 6 Decoding action and outcome combinations with striatal activity. Confusion matrices for striosomal (**a**) and matrix (**b**) decoding of action – reward outcome – puff outcome combinations using single session models.



Supplementary Fig. 7 Decoding future behavior with striatal activity. a, b Decoding of future switch/stay behavior and reward and puff outcome in striosomes (**a**) and matrix (**b**) using pseudo trials. **c** Accuracy of decoding left/right choices based on ITI activity in the 2 s preceding trial onset (*p < 0.05). Decoding accuracy in models based on single sessions was slightly better in a model using all neurons than only matrix neurons (mean ± SEM, p = 0.033; t = 2.00, two-sided repeated measures t-test, df = 12, n = 13, 100 pseudo trials). Source data are provided as a Source Data file.