Title

Combination of successor representation-based appetitive learning and individual representation-based aversive learning is a good strategy and can be implemented in the cortico-basal ganglia circuits

Short title

Opponent learning with different representations

Kenji Morita[1,2*], Kanji Shimomura[1,3], & Yasuo Kawaguchi[4,5]

[1] Physical and Health Education, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan

[2] International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo, Tokyo 113-0033, Japan

[3] Department of Behavioral Medicine, National Institute of Mental Health, National Center of Neurology and Psychiatry, Kodaira 187-8551, Japan

[4] Brain Science Institute, Tamagawa University, Machida 194-8610, Japan

[5] National Institute for Physiological Sciences (NIPS), Okazaki 444-8787, Japan

**Corresponding author*

Kenji Morita, Ph.D. (ORCID iD: orcid.org/0000-0003-2192-4248)

Physical and Health Education, Graduate School of Education, The University of Tokyo

7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

morita@p.u-tokyo.ac.jp

Author contributions

KM conceived of the possibility of differential representations in different cortico-basal ganglia pathways, and elaborated the idea in discussion with YK. KS considered to examine differential learning rates from positive and negative feedbacks, in discussion with KM. Based on these, KM developed the model presented in this manuscript, and conducted the simulations and analyses. KM and KS validated the program codes. KM drafted the manuscript, and revised it with inputs from KS and YK. According to the CRediT Taxonomy, Conceptualization: KM, KS, YK; Investigation: KM; Validation: KM, KS; Funding Acquisition: KM, YK; Writing – Original Draft Preparation: KM; Writing – Review & Editing: KM, KS, YK.

Conflict of interest

None

Abstract

While positive reward prediction errors (RPEs) and negative RPEs have equal impacts in the standard reinforcement learning, the brain appears to have distinct neural pathways for learning mainly from either positive or negative feedbacks, such as the direct and indirect pathways of the basal ganglia (BG). Given that distinct pathways may unevenly receive inputs from different neural populations and/or regions, how states or actions are represented can differ between the pathways. We explored whether combined use of different representations, coupled with different learning rates from positive and negative RPEs, has computational benefits. We considered an agent equipped with two learning systems, each of which adopted individual representation (IR) or successor representation (SR) of states. With varying the combination of IR or SR and also the learning rates from positive and negative RPEs in each system, we examined how the agent performed in a certain dynamic reward environment, specifically, in a simulated reward navigation task in a two-dimensional grid space, in which reward location changed over time. We found that combination of an SR-based system learning mainly from positive RPEs (i.e., appetitive learner) and an IR-based system learning mainly from negative RPEs (aversive learner) could achieve good performance, as compared to other combinations, including the ones where the two systems had the same ratios of positive- and negative-RPE-based learning rates and the ones where the two systems used the same representation, either IR or SR. In the combination of appetitive SR- and aversive IR-based systems, both systems show activities of comparable magnitudes with opposite signs, consistent with suggested profiles of BG pathways. These results suggest that particularly combining different representations with appetitive and aversive learning could be an effective learning strategy in certain dynamic reward environments, and it might actually be implemented in the cortico-BG circuits.

Introduction

In the standard reinforcement learning (RL), updates based on positive reward prediction errors (RPEs) and those based on negative RPEs are executed in the same manner. However, in the brain, there appear to exist distinct neural circuits that mainly conduct either learning from positive feedbacks or learning from negative feedbacks. Specifically, a number of findings have suggested or implied that the direct and indirect pathways of the basal ganglia (BG), originating from the striatal projection neurons (SPNs) expressing D1- and D2-type dopamine (DA) receptors (D1Rs and D2Rs), implement learning based primarily on positive and negative feedbacks, respectively [1-5]. There are also studies suggesting that distinct circuits involving the orbitofrontal cortex (OFC) operate for appetitive and aversive feedback-based learning [6]. Computational works suggest that dual learning systems can realize estimation of costs and benefits [7, 8], as well as estimation of not only the mean but also the uncertainty of rewards [9].

These existing dual learning-system models assume that both systems use the same way of representation of states or actions. Theoretically, various ways of representation can be considered, and different brain regions or neural populations may generally use different representations [10-12]. While there is evidence that the two BG pathways receive inputs from same types of corticostriatal neurons [13, 14], it has also been suggested that different neuron types [15-17] and/or cortical areas [18] may not evenly target these pathways. As for the suggested distinct OFC circuits for appetitive and aversive learning, the suggested circuits are the amygdala-OFC and OFC-nucleus accumbens (NAc) pathways, respectively [6]. Therefore, in both cases, it is conceivable that, between the appetitive and aversive learning systems, states/actions are represented by at least partially different neural regions or populations, and thus in different styles.

There are two largely different ways of state/action representation [19]. One is to represent each individual state/action separately. This simplest representation (or equivalent ones) has been explicitly or implicitly assumed to exist in the brain RL circuits in many previous studies. The other is to represent each state/action by a set of features (e.g., represent a point in a two-dimensional space by a set of x- and y-coordinates). Among various ways of feature-based representations, recent studies [20-24] suggest that representation of states by their successors, named the successor representation (SR) [25], may be adopted in the brain. SR contains information about state transitions in the environment under a given policy, and thereby enables the agent to quickly adapt to changes in distant outcomes through temporal difference (TD) RPE-based learning, without using an explicit model of the environment. It can thus beautifully explain [20] the empirical suggestions that both (apparently) model-based behavior and model-free or habitual behavior are controlled by the DA-cortico-BG systems although different portions appear to be responsible [26, 27] by assuming that SR and individual (punctate) representation (IR) are respectively used.

Given these suggestions and considerations, it seems possible that there exist two neural systems, which may differently learn from positive and negative RPEs, and may adopt different ways of state representations. In the present study, we examined its possible consequences through simulations. Specifically, we considered an agent equipped with two learning systems, and simulated its behavior in a reward navigation task in a certain dynamic reward environment, with the adopted representation, SR or IR, and the ratio of learning rates for positive versus negative TD-RPEs in each system freely and independently varied. We found that a particular combination of state representations and learning rate ratios, specifically, a combination of an SR-based system learning mainly from positive RPEs and an IR-based system learning mainly from negative RPEs could achieve good performance, as compared to other combinations.

Results

Performance of combined SR-based and IR-based systems, with learning rate ratio varied

We simulated a reward navigation task in a two-dimensional grid space, in which reward location changed over time (Figure 1), and examined the performance of an agent consisting of two systems, which adopted SR or IR and may have different ratios of learning rates for positive versus negative TD-RPEs; the two systems developed system-specific state values, and their average (named the integrated state value) was used for action selection and TD-RPE calculation (Figure 2) (see the Methods for details). We first examined the case in which one system employed the SR whereas the other adopted the IR, systematically varying the ratio of learning rates from positive and negative TD-RPEs for each system, denoted as α_{SR+} / α_{SR-} for the SR-based system and α_{IR+} / α_{IR-} for the IR-based system. The sums of the learning rates from positive and negative TD-RPEs (i.e., $\alpha_{SR+} + \alpha_{SR-}$ and $\alpha_{IR+} + \alpha_{IR-}$) were kept constant at 1, and the inverse temperature and the time discount factor were also kept constant at $\beta = 5$ and $\gamma = 0.7$. We counted how many rewards the agent obtained during the reward navigation task as a measure of performance, and examined how the obtained rewards, averaged across 100 simulations for each condition, varied depending on the learning rate ratios in both systems.

Figure 3A shows the results. As shown in the figure, the performance greatly varied depending on the conditions, and the best performance was achieved in the conditions in which α_{SR+} / α_{SR-} was high (much larger than 1) and α_{IR+} / α_{IR-} was low (much smaller than 1), i.e., the SR-based system learned mainly from positive TD-RPEs whereas the IR-based system learned mainly from negative TD-RPEs. Figure 3B shows the standard deviation (SD) and standard error of the mean (SEM) for the conditions where α_{SR+} / α_{SR-} times α_{IR+} / α_{IR-} was equal to 1 (i.e., the conditions on the horizontal diagonal in Figure 3A), and Figure 3C shows the frequency (number of times) that each combination of α_{SR+} / α_{SR-} and α_{IR+} / α_{IR-} gave the best mean performance over 100 simulations when 100 simulations were executed 50 times (including the one shown in Figure 3A,B) for each combination. As shown in these figures, the peak of the performance was rather broad, but was mostly in the range where α_{SR+} / α_{SR-} was large and α_{IR+} / α_{IR-} was small. We also examined the cases where the sums of the learning rates from positive and negative TD-RPEs were increased (1.25) or decreased (0.75), and also the inverse temperature and the time discount factor were varied ($\beta = 5$ or 10 and $\gamma = 0.7$ or 0.8). As shown in Figure 4, the result that combination of the SR-based system learning mainly from positive TD-RPEs and the IR-based system learning mainly from negative TD-RPEs achieved the best performance was preserved across these parameter changes.

Learning profiles of the model with different combinations of the learning rate ratios

Going back to the case with original parameter values used in Figure 3 ($\alpha_+ + \alpha_- = 1$, $\beta = 5$,

and γ = 0.7), we examined the learning curves of the model in the case where the ratio of positive versus negative TD-RPE-based learning rate was respectively high (4) and low (0.25) in the SR-based and IR-based systems (i.e., the case achieving the best performance), compared with the cases where the ratio was 1 in both systems or the ratio was low (0.25) and high (4) in the SR-based and IR-based systems. Figure 5A shows the mean learning curves in each case, i.e., the mean time (number of time steps) used for obtaining the first, second, third, .. reward *placed* in each rewarded epoch, averaged across eight (from the second to the ninth) rewarded epochs and also across simulations (if a reward was placed in the first epoch and obtained in the second epoch, for example, it was regarded as the last reward in the first epoch rather than the first reward in the second epoch). Only the cases in which reward was obtained in not smaller than a quarter of (25 out of total 100) simulations were plotted, in order to see the general average behavior of the model. Comparing the case with high α_{SR+} / α_{SR-} & low α_{IR+} / α_{IR-} (brown line) and the case with low α_{SR+} / α_{SR-} & high α_{IR+} / α_{IR-} (blue-green line), the time consumed for obtaining the first reward in an epoch, as well as the asymptotic value of the time spent for reward acquisition, were much longer in the latter case. The case with equal ratios for both systems (red line) showed an intermediate behavior.

We also looked at how each system developed system-specific values of the states, again in the three cases where α_+ / α_- for the SR-based and IR-based systems were 4 and 0.25, 1 and 1, or 0.25 and 4. Specifically, we looked at single-simulation examples of those values just after the agent obtained the last reward in the last (9th) rewarded epoch in single simulations, in which that reward was placed at the "special candidate state (with high probability of becoming a rewarded state: see the Methods)" in that epoch, in the three cases. Figure 5Ba shows the best-performance case where α_{SR+} / α_{SR-} was high (4) and α_{IR+} / α_{IR-} was low (0.25). The left panels show both the system-specific and integrated values, and the middle and right panels show only the integrated values in different ways. As shown in the figure, the system-specific state values of the SR-based system (red line in Figure 5Ba left) were all positive whereas those of the IR-based system (blue line) were all negative, and these two look largely symmetric. Meanwhile, the average of these two system-specific values, i.e., the integrated state values (black lines in Figure 5Ba left and middle; also shown as a heat map in Figure 5Ba right), have much smaller magnitudes but show a clear pattern, which is expected to approximate the true state values under the policy that the agent was taking. Figure 5Bb shows the case where both α_{SR+} / α_{SR-} and α_{IR+} / α_{IR-} were 1. In this case, both systems developed similar values. Figure 5Bc shows the case where α_{SR+} / α_{SR-} was low (0.25) and α_{IR+} / α_{IR-} was high (4). In this case, the SR- and IR-system-specific values were negative and positive, respectively, in contrast to the case of Figure 5Ba with high α_{SR+} / α_{SR-} and low α_{IR+} / α_{IR-}. The integrated values have comparable magnitudes to the other two cases, but were negative in many states. Comparing the right panels of Figure 5Ba and Figure 5Bc, the peak of the value function looks sharper in the latter, presumably reflecting the smaller degree of positive TD-RPE-dependent learning in the SR-based system.

Cases where both systems employ only the IR or only the SR

We also examined the cases where both systems employed only the SR, or only the IR. Figure 6A and 6B show the performance results for the SR only and IR only cases, respectively, for the same sets of parameters as shown in Figure 4 for the case of SR and IR combination. Notably, if the ratio of positive versus negative TD-RPE-based learning rates was equal between both systems that employ only the SR or IR, the two systems behaved in exactly the same manner, and thus such conditions (on the vertical lines in Figure 6A and 6B) were equivalent to having only a single system. Also, Figure 6A and 6B only show either the left or right side, because "$\alpha_{1+} / \alpha_{1-} = 0.2$ and $\alpha_{2+} / \alpha_{2-} = 3$", for example, are equivalent to "$\alpha_{1+} / \alpha_{1-} = 3$ and $\alpha_{2+} / \alpha_{2-} = 0.2$" given that both systems 1 and 2 employed the same representation (SR or IR) so that we examined only one of these, and Figure 6A shows the left side whereas Figure 6B shows the right side just because such arrangements might facilitate visual comparisons with Figure 4, where the ratios for the SR-based system and the IR-based system were plotted on the axes rising to the left and the right, respectively.

Comparing Figure 6A and 6B, SR-SR combination generally achieved better performance than IR-IR combination. This may not be surprising, given that a known feature of SR-based learning is sensitive adaptation to changes in rewards in the environments [20, 21], which occurred in our simulated task. Besides, during the initial no-reward epoch, the SR-based system could acquire, through TD learning of SR features, an SR under the random policy, which presumably acted as beneficial "latent learning" [25], while the IR-based system could not learn anything. Likewise, although the location of special reward candidate state and thus the optimal policy varied from epoch to epoch and SR is a policy-dependent representation, SR features acquired and updated in the agent should contain information about the basic structure of the grid space, which could help reward navigation throughout the task. Next, whether having two systems with different ratios of positive and negative TD-RPE-based learning rates was advantageous, in terms of the mean performance in this task, appears to be not clear in the cases where both systems employed only the SR or IR than in the case of the combination of SR- and IR-based systems. Last and most strikingly, the best combination of the SR-based and IR-based systems, i.e., the combination of the SR-based system learning mainly from positive TD-RPEs and the IR-based system learning mainly from negative TD-RPEs outperformed any combinations of two SR-based systems or two IR-based systems for all the sets of parameters ($\alpha_+ + \alpha_-$, β, and γ) that were so far examined (compare Figure 4 and Figure 6A,B; also shown in Figure 6C).

More comprehensive examination of the parameter space

So far we examined the cases where the sum of the learning rates from positive and negative TD-RPEs ($\alpha_+ + \alpha_-$) in both systems was one of the three values (0.75, 1, or 1.25) and the learning rate for the update of SR features was fixed at 0.05. In order to examine the parameter space more widely,

we varied the learning rate from positive or negative TD-RPEs in each of the two systems freely from 0.2, 0.35, 0.5, 0.65, or 0.8 and the learning rate for the update of SR features from 0.05, 0.1, 0.15, 0.2, or 0.25 (only for the cases where at least one of the two systems employed SR), also varying the time discount factor (γ = 0.6, 0.7, or 0.8) and the degree of choice exploration-exploitation (i.e., inverse temperature) (β = 5, 10, or 15), for each of the cases where the two systems employed SR and IR, SR only, or IR only. Figure 7 shows the mean performance of the model consisting of SR- and IR-based systems. Each row of panels in Figure 7 shows the mean performance for each set of time discount factor and inverse temperature (shown in the left), varying the learning rate for the update of SR features (shown in the top), projected onto the plane consisting of the ratios of the learning rates from positive and negative TD-RPEs in the two systems ($\alpha_{SR+}/\alpha_{SR-}$ and $\alpha_{IR+}/\alpha_{IR-}$); there were 25 cases with $\alpha_{SR+}/\alpha_{SR-} = \alpha_{IR+}/\alpha_{IR-} = 1$, which were difficult to draw on the single point (1, 1) in this projected plane and thus omitted. Figure 8 shows the mean performance of the model consisting of two SR-based systems or two IR-based systems, projected onto the plane consisting of α_{1+}/α_{1-} and α_{2+}/α_{2-}; 25 cases with $\alpha_{1+}/\alpha_{1-} = \alpha_{2+}/\alpha_{2-} = 1$ were omitted for the same reason as above. Table 1 shows the sets of learning rate parameters that gave top ten mean performance for each set of time discount factor and inverse temperature (shown in the left) in the model consisting of SR- and IR-based systems (left), two SR-based systems (middle), and two IR-based systems (right). For the model consisting of SR- and IR-based systems, combinations of appetitive SR-based system ($\alpha_+/\alpha_- > 1$) and aversive IR-based system ($\alpha_+/\alpha_- < 1$) are shown in red.

Regarding the model consisting of SR- and IR-based systems (Figure 7 and Table 1), under the examined conditions differing in the degrees of temporal discounting and/or choice exploration-exploitation (each row of panels in Figure 7 and sub-tables in Table 1), combination of appetitive SR-based and aversive IR-based systems (shown in red in Table 1) with relatively small learning rate for SR feature update ($\alpha_{SRfeature}$) generally achieved good performance. As $\alpha_{SRfeature}$ increased, combination of the α_+/α_- ratios in the two systems that achieved good performance approached to the vertical diagonal line (i.e., similar α_+/α_- ratios in both systems), but the achieved good performance itself tended to decrease when $\alpha_{SRfeature}$ exceeded 0.1 or 0.15.

Regarding the models where both systems employed only SR or IR (Figure 8 and Table 1), the SR-only models generally outperformed the IR-only models. For the SR-only model, increase in $\alpha_{SRfeature}$ up to 0.15 or 0.2 improved the performance in many cases, but further increase appears to be not generally beneficial. As for the SR-only model, the case with (γ, β) = (0.6, 15), performed well, comparably to the case with (γ, β) = (0.7, 10). So we further examined cases with (γ, β) = (0.5, 15), (0.5, 20), and (0.6, 20). Table 2 shows the sets of learning rates giving top ten mean performance for these three cases, which look largely comparable to the cases (γ, β) = (0.6, 15) or (0.7, 10). As for the IR-only model, the cases with small inverse temperature (β = 5) performed well. So we further examined cases with β = 2.5, but performance was generally worse than the cases with β = 5 (best mean performance was 107.2, 107.76, and 105.12 in the cases with γ = 0.6, 0.7, and 0.8, respectively).

Comparing the three types of models using SR and IR, SR only, or IR only, under each of the nine (3×3) sets of time discount factor and inverse temperature (γ = 0.6, 0.7, 0.8 and β = 5, 10, 15; shown in Figures 7 and 8 and Table 1) that were examined for all the three types of models, best mean performance, among the examined sets of learning rate parameters, was achieved by a combination of SR- and IR-based systems. Under seven out of the nine sets, combination of appetitive SR- and aversive IR-based systems achieved the best performance, and even under the remaining two, best performance was achieved by a combination where the ratio of learning rates from positive over negative TD-RPEs was higher in the SR-based system than in the IR-based system.

Dependence on task properties

So far we examined the agent's performance in the particular task. We examined how it could differ if task properties change. The original task contained a stochastic nature in a sense that reward was placed at a particular state (named the special reward candidate state) with a high probability (60%) in a given epoch but with the remaining probability reward was placed at one of the other (normal) candidate states. We examined what happened when this stochasticity was reduced or removed in the case of the model consisting of SR-based and IR-based systems with the original set of parameters used in Figures 3 and 5 ($\alpha_{SR+} + \alpha_{SR-} = \alpha_{IR+} + \alpha_{IR-} = 1$, $\beta = 5$, and $\gamma = 0.7$). Figure 9A shows the results for the cases where the probability of reward placement at the special candidate state was varied from 70% to 100%. As shown in the figure, as the stochasticity reduced, combinations of α_{SR+} / α_{SR-} and α_{IR+} / α_{IR-} that gave good performance gradually shifted, and when the stochasticity was totally removed, best performance was achieved when both α_{SR+} / α_{SR-} and α_{IR+} / α_{IR-} were high values.

We also examined the effects of another variation in the task properties. In the original task, new reward was not placed until the agent obtained the previous reward. This property, coupled with the stochastic placement of reward, is considered to make perseveration-like behavior quite maladaptive and increase the importance of learning from negative feedbacks. We examined what happened if this property was weakened by introducing periodic resets of reward placement into the original task, again in the case of the model consisting of SR-based and IR-based systems with the original set of parameters. Figure 9B shows the results for the cases where reward location (state) was reset at every 500, 250, 100, or 50 time steps in the rewarded epochs; at the reset timings, reward was placed at the special candidate state for the epoch with 60% and placed at one of the other (normal) candidate states with 5% each. As shown in the figure, as the resets became more frequent, combinations of α_{SR+} / α_{SR-} and α_{IR+} / α_{IR-} that gave good performance again gradually shifted from high α_{SR+} / α_{SR-} and low α_{IR+} / $\alpha_{IR-}$to high α_{SR+} / α_{SR-} and also high α_{IR+} / α_{IR-}, while overall sensitivity to α_{IR+} / α_{IR-} looked diminished.

We further examined three different variations of the original task. The first variation was

abolishment of multiple rewarded epochs with different placements of special reward candidate state. Specifically, we examined the case where there was only a single rewarded epoch with 4500 time steps, instead of nine epochs with 500 time steps for each, following the initial 500 time-steps no-reward epoch. The special candidate state was varied only across simulations. The second variation was removal of the stochasticity of reward placement within each epoch, together with shortening of the duration of each rewarded epoch (from 500 time steps to 100 time steps) and increasing the number of rewarded epochs (from 9 to 45). The third variation was abolishment of special reward candidate state. Specifically, we came back to the original nine 500 time-steps rewarded epochs, but the probability of reward placement at the special candidate state was changed from the original 60% to 1/9 (11.11..%) so that reward was placed at each of the nine candidate states with equal probability (1/9). Figures 9C-E show the performance of the model consisting of SR-based and IR-based systems with the original set of parameters in these three variations. As shown in the figures, the patterns look similar to the case of the original task (Figure 3).

It was rather surprising that similar pattern appeared even in the last case, because reward was placed at any candidate states with equal probabilities and so it was no longer a task where the agent could learn where the special candidate state was in each epoch. However, the nine candidate states were neighboring with each other on the two edges of the grid space, and thus it would still have been the case that next reward was placed at a state near from the previously rewarded state with a relatively high probability and the agent could exploit it through positive TD-RPE-based learning. As a control, we examined a task in which positive TD-RPE-based learning was expected to have little meaning. Specifically, after 500 time-steps no-reward epoch, rewarded state was determined by a fixed order, namely, (5, 1), (5, 5), (1, 5), and again (5, 1), and these were repeated throughout the task. Figure 9F shows the performance of the model consisting of SR-based and IR-based systems with the original parameters. As shown in the figure, best performance, though rather low as expected, was achieved when both α_{SR+} / α_{SR-} and α_{IR+} / α_{IR-} were low values. This is reasonable because only quick erasing of the memory of previous reward is considered to be adaptive in this control task.

These results together indicate that the combination of appetitive SR and aversive IR systems does not always achieve the best performance, but does so in certain environments where reward placement is dynamically changing but still learnable using both positive and negative TD-RPEs.

Performance of the two-system models in the two-stage tasks

Whether humans or animals take model-based/goal-directed or model-free behavior has been widely examined by using the so-called two-stage (or two-step) tasks [28, 29]. Therefore we also examined how our models consisting of two systems performed in such tasks. We simulated the two-stage task [28]. At the first stage, there were two choice options. Selection of each of these two options lead to one of two pairs of second-stage options with fixed probabilities (70% or 30%) (Figure 10A).

Then, selection of one of the second-stage options lead to reward or no-reward outcome. The probability of reward for each second-stage option was independently set according to Gaussian random walk with reflecting boundaries at 0.25 and 0.75 (an example is shown in Figure 10Ba). The IR-based system learned the value of each first- and second-stage option through RPE-based updates. The SR-based system had the SR of the first- and second-stage options, and leaned their values through RPE-based updates of the weights of the approximate value function. The initial values of the SR matrix were set incorporating the probabilities of the transitions from the first stage to the second stage and also assuming random choice policy at the second stage, and the SR matrix was updated by using the prediction error of SR features.

Figure 10Ca shows the performance of the model with ($\alpha_+ + \alpha_- = 1$, $\beta = 5$, and $\gamma = 1$) and the ratio of the learning rates for positive and negative RPEs in each system ($\alpha_{SR+}/\alpha_{SR-}$ and $\alpha_{IR+}/\alpha_{IR-}$) varied. As shown in the figure, combination of appetitive SR and aversive IR systems did not give a good performance in this task. We also examined a variant of the task, where the probabilities of reward for the second-stage options were set to specific values, which changed three times during the task (Figure 10Bb), so that differences among the values and their temporal changes could be clearer than the original case using random walk. Figure 10Cb shows the performance of the same model, and here again combination of appetitive SR and aversive IR systems did not give a good performance.

A possible reason why the good performance of such a combination observed in the navigation tasks was not generalized to the two-stage tasks was that the two-stage tasks imposed only selection from two options at both stages while the navigation tasks imposed selection from more than two options. In the case with only two choice options, an increase in the choice probability of one option simultaneously means the same magnitude of decrease in the choice probability of another option. Intuitively, generalization of learning from positive feedbacks for one option appears to have limited merit if choice probabilities are determined by the value difference between two options and there is no other option to which generalization can be applied. Therefore we simulated a variant of two-stage task, in which there were three, rather than two, first-stage options and three pairs of second-stage options (Figure 10D). Figure 10E shows the performance of the same model in this task. As shown in the figure, in this case, combination of appetitive SR-based system and aversive IR-based system achieved a relatively good performance. Conditions with similar intermediate α_+/α_- ratios for both systems, or even with higher α_+/α_- ratio for the IR system than for the SR system, achieved a good, or even better, mean performance, but shifts from such conditions towards appetitive IR- and aversive SR-based systems resulted in a rapid decrease in the performance while shifts towards appetitive SR- and aversive IR-based systems resulted in a milder decrease in the performance (although the overall range of the changes in the mean performance was rather small). Given this, we considered that the good performance of the combination of appetitive SR- and aversive IR-based systems observed in the navigation tasks could be generalized to the two-stage task with more than two options at least to a certain extent.

Figure 10F and G show the performance of the models in which both of the two systems employed only SR or IR, respectively. The ranges of the performance of these two types of models look comparable with each other, and also comparable to the performance range of the model consisting of SR- and IR-based systems. Therefore, the superiority of the solely SR-based model over the solely IR-based model, and also the superiority of the combined SR- and IR-based systems over the solely SR- (or IR-) based system, observed in the navigation tasks was not generalized to this two-stage task. It has been pointed out [30] that model-based control actually provides little performance benefits in the original two-stage task [28] as well as in several variants. Although model-based control can still be beneficial in certain two-stage tasks [30], advantages of model-based or SR-based control, and potentially also of the combined SR- and IR-based systems, might appear more clearly in tasks with more than two stages such as the navigation tasks.

Discussion

We found that combination of SR-based system learning mainly from positive TD-RPEs and IR-based system learning mainly from negative TD-RPEs showed superior performance in certain dynamic reward environments. Below we discuss issues related to the results of our simulations and the assumptions of our model from computational and neurobiological perspectives.

Reasons for the superior performance of the combined appetitive SR- and aversive IR-based systems

As possible reasons why the combination of appetitive SR- and aversive IR-based systems performed well, the following two are considered. The first possible reason comes from an asymmetry between positive- and negative-error-based learning. When the agent unexpectedly encountered a reward at a certain state, next reward was placed at the same state or nearby states with a certain high probability (given that there was learnable structure). Then, quickly revising up the value of not only the rewarded state itself but also any states from which the rewarded state could be easily reached (i.e., states having high SR feature values) would be beneficial. In other words, SR-dependent generalization of positive TD-RPE-based value updates would be beneficial. The result that the combination of appetitive SR- and aversive IR-based systems performed relatively well in the two-stage task with three first-stage options but not in the task with two options seems in line with this, because a key difference between the task variant with two options only and the one with three options would be that generalization would have a larger merit in the latter than in the former. Next, assume that the agent obtained rewards several times at the same or nearby states (in the navigation task), and then reward was placed at a distant state. This time, the agent was likely to unexpectedly encounter reward omission. Critically, through the repetitive acquisition of rewards at the nearby states, the agent's policy was likely to be sharpened and therefore revising down the values of only the states right on the sharpened policy would already be effective. In other words, as for negative TD-RPE-based value updates, IR-based narrow effect would be largely sufficient, and too much SR-dependent generalization could rather be harmful.

The second possible reason for the good performance of the combination of appetitive SR- and aversive IR-based systems, which is considered to be also related to the last point of the above, comes from the policy-dependence of SR. SR reflects state transitions under the policy that has been used. Thus, when reward is relocated at a distant place and thereby optimal policy is drastically changed, SR under the policy optimized for the previous reward should significantly differ from the one under the new optimal policy, and SR becomes updated as the agent changes its policy. In fact, through TD-RPE-based learning, in contrast to direct calculation of state values by multiplication of SR and reward at each state (if it could be estimated), value function after reward relocation can in principle be approximated even with SR under the previously near-optimal policy, as well as with

arbitrary state representation, as suggested in [20]. Nonetheless, SR under the previously near-optimal policy may not have fine information about state transitions near the new reward location, implying potential difficulty in learning from negative TD-RPEs for SR-based system, which could contribute to the superiority of the combination of appetitive SR- and aversive IR-based systems. Difficulty in the case with drastic change in the goal location and the optimal policy has previously been shown [31] for learning using successor features, which are generalization of SR [32]. Increasing the learning rate for SR feature update ($\alpha_{\mathrm{SRfeature}}$) could potentially mitigate this issue. In our simulations with $\alpha_{\mathrm{SRfeature}}$ varied (Figure 7 and Table 1), as $\alpha_{\mathrm{SRfeature}}$ increased, combination of the α_{+}/α_{-} ratios in the two systems that achieved good performance approached to the ones with similar ratios in both systems, in line with the above argument. At the same time, however, increase in $\alpha_{\mathrm{SRfeature}}$ did not generally improve the highest achievable performance of the model with SR- and IR-based systems; increase up to 0.1 improved it in some but not other conditions, and further increase could rather worsen the highest achievable performance presumably because state representation became too unstable, while increase in $\alpha_{\mathrm{SRfeature}}$ up to 0.15 or 2 was beneficial in many cases for the model consisting of two SR-based systems (Figure 8 and Table 1). From the results that relatively small $\alpha_{\mathrm{SRfeature}}$ was good for the model with appetitive SR- and aversive IR-based systems, we speculate that policy-independent state representation based on the transition structure, specifically, default representation [33], could alternatively be used.

Appetitive and aversive learning in distinct neural pathways

There have been a number of empirical studies suggesting that the D1/direct and D2/indirect pathways of BG are primarily involved in appetitive and aversive learning, respectively. Parkinson disease patients on and off medication showed opposite patterns of learning, better from positive and negative feedbacks, respectively, presumably through changes in the balance of D1/direct and D2/indirect pathways [1]. Blockade of the direct or indirect pathway impaired reward or avoidance learning, respectively [2]. Optogenetic stimulation of D1 or D2 SPNs induced reinforcement or punishment [3], and mimicked increase or decrease in action values [34], respectively. Neurons in the globus pallidus externus on the indirect pathway showed inhibitory response to valueless objects [35]. The direct and indirect pathway SPNs encoded reward and no-reward outcomes, respectively [36]. The direct and D2/indirect SPNs were shown to be essential for initial learning of instrumental contingency and flexible update of behavior after contingency change, respectively [37, 38].

At the synaptic level, D1R- and D2R-mediated plasticities have opposite DA dependences [39]. Spines on D1 and D2 SPNs were enlarged by phasic DA rise [40] and dip [4], respectively, both depending on protein kinase A (PKA). Recent work [5] further demonstrated that phasic activation of DA neurons, inducing DA rise, increased net activity of PKA in D1 SPNs while minimally modulating it in D2 SPNs (in their Fig. 4b-d), whereas phasic inactivation of DA neurons, inducing DA dip, increased net PKA activity in D2 SPNs but not significantly in D1 SPNs (in their Fig. 4e-g). This

suggests that PKA-dependent plasticities in D1 and D2 SPNs are not just in the opposite directions but have differential sensitivities to DA rise and dip, implying that the D1 and D2 pathways primarily learn from positive and negative TD-RPEs, respectively, although PKA-independent plasticity processes may also exist.

Theoretical studies explored potential functional merits of having such distinct BG pathways for appetitive and aversive learning. A study [7] suggested that the D1/direct and D2/indirect pathways could learn the benefits and costs of the same action, respectively, and DA-dependent differential weighting of costs and benefits could be achieved. Other studies [9] suggested that these pathways could learn not only the mean but also the uncertainty of reward, and the set of plasticity rules achieving it could also realize estimation of costs and payoffs [8].

Differences in the inputs to the neural pathways for appetitive and aversive learning

These existing models assumed that the D1/direct and D2/indirect pathways employ the same way of representation of states or actions, specifically, the IR. Since different neural regions or populations may generally adopt different ways of representation, as in the case of egocentric and allocentric spatial representations [10-12], the two BG pathways might also use different representations if their input sources are not identical.

There are two major types of cortical pyramidal cells projecting to the striatum, namely, intratelencephalic (IT) neurons, which project to bilateral striatum, and pyramidal-tract (PT) neurons, which target ipsilateral striatum [16, 41]. Electrophysiological and optogenetic studies have shown that both D1/direct and D2/indirect pathways receive inputs from both IT and PT neurons [13, 14]. However, anatomical studies suggested that D1 and D2 SPNs predominantly receive inputs from IT and PT neurons, respectively [15, 16]. Model fitting of previously published paired-pulse ratio data also suggested IT→D1 and PT→D2 preferential activation through short-term synaptic plasticity [17]. At a larger spatial scale, it has been shown that different cortical regions innervate the D1/direct and D2/indirect pathways in different proportions [18]. In particular, the somatosensory and limbic cortical regions preferentially innervated the direct pathway whereas the primary motor cortex preferentially targeted the indirect pathway.

Given these, it is conceivable that the D1/direct and D2/indirect pathways unevenly receive inputs from different sources, which could represent states in different ways. Other than the BG pathways, the amygdala-OFC and OFC-NAc pathways have also been suggested to be specifically involved in appetitive and aversive learning, respectively [6]. In this case, the pathways have different input structures (i.e., amygdala and OFC), which may use different state representations.

What types of state representation are used in the different input neuron types or regions?

Given that the D1/direct and D2/indirect pathways are unevenly innervated by the IT and PT cortical neurons and/or by the sensory/limbic and motor cortices as suggested, an emerging question is whether either type of state representation, in particular, SR or IR, is preferentially used in the IT and PT neurons or in the sensory/limbic and motor cortices.

As for neuron types, recent studies have suggested that sensitivity to outcome devaluation, which is indicative of model-based/goal-directed behavior, is mediated via IT, rather than PT, neurons projecting from the prelimbic cortex to the posterior dorsomedial striatum [42, 43]. Since SR-based TD learning also enables sensitive adaptation to changes in outcome values [20, 21], these results would also be consistent with the possibility that IT neurons, but not PT neurons, are engaged in SR-like representation, while the results also suggest that SR-based learning should occur from negative feedbacks (i.e., devaluation) at least to some extent. Local connections between IT neurons are less reciprocal (i.e., more unidirectional) than those between PT neurons [44], and axonal projections of individual IT neurons to the striatum tend to have more widespread endpoints than those of PT neurons [45]. These properties might also be in line with the possibility that IT neurons have SR-like state representation, because SR is based on the directional transition relationships between states and also calculation of state value using SR requires access to not only that state but also all the states which are reached from that state through transitions. Regarding cortical regions, BOLD signals consistent with SR have been observed in hippocampal–entorhinal cortex [22], and recent work found evidence for SR in decoded activity in sensory cortex [24].

Given that IT neurons and/or sensory/limbic cortices preferentially innervate the BG direct pathway while PT neurons and/or motor cortex preferentially innervate the indirect pathway as mentioned above, the suggested possible employment of SR in IT neurons or in the hippocampal–entorhinal or sensory cortices indicates that the direct pathway preferentially employs the SR while the indirect pathway does not (Figure 11). Intriguingly, this appears to match the best-performance combination in our simulations: SR-based and IR-based systems learning mainly from positive and negative TD-RPEs, respectively. It is tempting to speculate that such a combination has been equipped in the brain through evolution because it is adaptive. In the best combination, the SR-based system-specific values became positive while the IR-based system specific values became negative, and these two looked largely symmetric (Figure 5Ba). If these systems are implemented by the BG direct and indirect pathways, respectively, such symmetry means that these pathways show largely similar activity patterns. This seems in line with the observation [46] that the striatal direct- and indirect-pathway neurons showed concurrent activation, rather than antagonistic activation that was expected from the classical view.

Strengths, limitations, and perspectives

As discussed above, our dual systems model with different representations coupled with

differential appetitive and aversive learning can potentially be well mapped onto the neuron type- and/or cortical area-specific circuitry in the cortico-BG system, going beyond the previous models that took into account the two BG pathways but not the diversity and complexity of the neocortex. Our model is still algorithmic, but it could be a step onto the anticipated bridge from algorithmic to implementational levels [47].

We have shown that combination of appetitive SR-based and aversive IR-based systems generally achieved good performance in several variations of reward navigation tasks, and that this tendency was also, to a certain extent, observed in a variant of two-stage task with three first-stage options. Sets of parameters that we examined, however, were still limited, and it is possible that other combinations with unexamined sets of parameters perform well. It also remains to be seen whether the superiority of the combination of appetitive SR-based and aversive IR-based systems is applied also to other tasks, including more complex tasks. As for biological plausibility, although there are findings that appear to be in line with the architecture of our model as discussed above, the links are still speculative. There are also numerous missing links. To mention one, we assumed that TD-RPEs are calculated from the integrated values, but how it can be achieved remains elusive. Previously two of us proposed that the IT→D1/direct and PT→D2/indirect pathways respectively represent the current/upcoming and previous values and they are used for TD-RPE calculation [48, 49], but we considered different possibilities about these pathways in the present work. Integrating our present model with mechanisms of TD error calculation [50] is an important future direction.

Our results suggest that combination of SR-based learning mainly from positive TD-RPEs and IR-based learning mainly from negative TD-RPEs is a good learning strategy. Whether humans or animals indeed take such a strategy can potentially be examined through behavioral experiments to examine whether the behavior exhibits features of SR-based learning (c.f., [21]) differently between the cases with positive and negative TD-RPEs. Brain substrates can also be explored by neuroimaging (c.f., [22, 24]).

Methods

Simulated reward navigation task

We simulated a reward navigation task in a certain dynamic reward environment. An agent was moving around in a 5 × 5 grid space (Figure 1A). The agent started from the fixed start state, which was location (1, 1), and moved to one of the neighboring states (two states at the corners, three states at the edges, and four states elsewhere) at each time step. During the initial 500 time steps, there was no reward, and the agent just moved around (Figure 1B). After that, a reward (always size 1) was introduced into one of the state (location) in the space. There were nine reward candidate states, where reward could potentially be placed, namely, (1, 5), (2, 5), (3, 5), (4, 5), (5, 1), (5, 2), (5, 3), (5, 4), and (5, 5) (i.e., the states on the two edges apart from the start state). During the next 500 time steps following the initial no-reward epoch, one of these nine candidate states was specified as the *special reward candidate state*, whereas the remaining eight candidate states were regarded as *normal candidate states*. There were in total nine rewarded epochs, each of which lasted for 500 time steps, and each one of the nine candidate states became the special candidate state in one of the nine epochs; the order was determined by pseudorandom permutation in each single simulation.

In the rewarded epochs, if the agent reached the rewarded state and obtained the reward, the agent was immediately carried back to the start state, and a new reward was introduced into a state, which was the special reward candidate state with 60% probability and one of the eight normal reward candidate states with an equal probability (i.e., 5%) each (Figure 1C).

We also examined variations of the task. In all the variations, there was the initial 500 time-steps no-reward epoch. In the first variation, the probability that reward was placed at the special candidate state, which was 60% in the original task, was varied to 70, 80, 90, or 100%. In the second variation, periodic resets of reward placement were introduced. Specifically, reward location (state) was reset at every 500, 250, 100, or 50 time steps in the rewarded epochs; at the reset timings, reward was placed at the special candidate state for the epoch with 60% and placed at one of the other (normal) candidate states with 5% each. In the third variation, the original nine rewarded epochs with different placements of special reward candidate state (500 time steps each) were abolished and instead there was only a single rewarded epoch with 4500 time steps. For this task variation, the special candidate state was varied only across simulations. In the fourth variation, reward was always placed at the special candidate state in each epoch and the duration of rewarded epoch was shortened from 500 time steps to 100 time steps while the number of rewarded epochs was increased from 9 to 45. The order of special reward candidate states was determined by five consecutive pseudorandom permutations of the nine candidate states. In the fifth variation, the probability of reward placement at the special candidate state was changed to 1/9 (11.11..%) so that reward was placed at each of the nine candidate states with equal probability (1/9). In the sixth variation, rewarded state was determined by a fixed order, namely,

(5, 1), (5, 5), (1, 5), and again (5, 1), and these were repeated throughout the task.

Model agent equipped with two learning systems

The agent's behavior was controlled by an RL model consisting of two learning systems (Figure 2). The two systems may use different ways of state representation. We considered the successor representation (SR) and the individual representation (IR) (also called the "punctate" representation), and examined the cases where system 1 and 2 used the SR and IR, respectively, and also the cases where only the SR or IR was used in both systems. Each system had its own *system-specific value* of each state (detailed below). The mean (average) of the system-specific values of the two systems was calculated and used as the *integrated value* of each state S, denoted as $V(S)$. At any states other than the rewarded state, the agent selected an action to move to one of the neighboring states depending on their integrated state values in a soft-max manner. Specifically, the agent selected the action to move to state S_i, among the neighboring states S_j, with probability

$$\exp(\beta V(S_i)) / \Sigma_j\{\exp(\beta V(S_j))\},$$

where β indicates the inverse temperature parameter, representing the degree of exploitation over exploration. When the agent was at state $S(t)$ at time-step t and moved to $S(t+1)$ at $t+1$, where $S(t)$ was not the rewarded state, TD-RPE was calculated based on the integrated state values:

$$\delta(t+1) = 0 + \gamma V(S(t+1)) - V(S(t)),$$

where γ indicates the time discount factor. When the agent was at $S(t)$, which was the rewarded state, TD-RPE

$$\delta(t+1) = 1 + 0 - V(S(t))$$

was calculated, again using the integrated state value.

In the system using the IR, the system-specific state value, $V_{system\text{-}i}(S(t))$, where i was 1 or 2, was directly updated based on the TD-RPE $\delta(t+1)$, with potentially different learning rates depending on whether the TD-RPE was positive (non-negative) or negative:

$$V_{system\text{-}i}(S(t)) \leftarrow V_{system\text{-}i}(S(t)) + \alpha_{i+}\delta(t+1) \quad \text{if } \delta(t+1) \geq 0, \text{ or}$$

$$V_{system\text{-}i}(S(t)) \leftarrow V_{system\text{-}i}(S(t)) + \alpha_{i-}\delta(t+1) \quad \text{if } \delta(t+1) < 0,$$

where α_{i+} and α_{i-} indicate the learning rates of the system i for positive (non-negative) and negative TD-RPEs, respectively, and they were systematically varied as described in the Results. The initial values of the system-specific state values of the system using the IR were set to 0.

In the system using the SR, the system-specific state value, $V_{system\text{-}i}(S(t))$, where i was 1 or 2, was calculated as a linear function of the feature variables of the state. In the SR, every state S_i was represented by a set of feature variables, which were estimated cumulative discounted future occupancies of all the states S_j (j = 1, ..., 25), denoted as $\sigma(S_i, S_j)$. These feature variables were updated through TD learning of SR features [51, 52]. Specifically, when the agent was at $S(t)$ at t and moved to $S(t+1)$ at $t+1$, where $S(t)$ was not the rewarded state, the TD errors for SR features were calculated

as:

$$\delta_{SRj}(t+1) = 1 + \gamma\sigma(S(t+1), S_j) - \sigma(S(t), S_j) \quad \text{if } S(t) \text{ was } S_j, \text{ or}$$

$$\delta_{SRj}(t+1) = 0 + \gamma\sigma(S(t+1), S_j) - \sigma(S(t), S_j) \quad \text{if } S(t) \text{ was not } S_j,$$

for all the states S_j (j = 1, ..., 25). If $S(t)$ was the rewarded state, the term $\gamma\sigma(S(t+1), S_j)$ was dropped. Using these TD errors, $\sigma(S(t), S_j)$ (j = 1, ..., 25) were updated as:

$$\sigma(S(t), S_j) \leftarrow \sigma(S(t), S_j) + \alpha_{SRfeature}\delta_{SRj}(t+1),$$

where $\alpha_{SRfeature}$ indicates the learning rate for the update of SR features and was set to 0.05 unless otherwise mentioned; $\alpha_{SRfeature}$ = 0.1, 0.15, 0.2, and 0.25 were also examined in Figures 7 and 8 / Tables 1 and 2. The initial values of all the feature variables were set to 0. The system-specific state value, $V_{system\text{-}i}(S(t))$, was given as a linear function of the feature variables as:

$$V_{system\text{-}i}(S(t)) = \Sigma_j\{w_j\sigma(S(t), S_j)\}.$$

The weights (coefficients) w_j (j = 1, ..., 25) were updated based on the TD-RPE $\delta(t+1)$ described above, with potentially different learning rates depending on whether the TD-RPE was positive (non-negative) or negative:

$$w_j \leftarrow w_j + \alpha_{i+}\sigma(S(t), S_j)\delta(t+1) \quad \text{if } \delta(t+1) \geq 0, \text{ or}$$

$$w_j \leftarrow w_j + \alpha_{i-}\sigma(S(t), S_j)\delta(t+1) \quad \text{if } \delta(t+1) < 0,$$

where α_{i+} and α_{i-} indicate the learning rates of the system i for positive (non-negative) and negative TD-RPEs, respectively, and they were systematically varied as described in the Results. The initial values of the weights w_j were set to 0.

Simulated two-stage tasks

We also simulated the two-stage task [28] and its variants. In our simulation of the original task (Figure 10A), there were two choice options at the first stage. Selection of one of them lead to each of two pairs of second-stage options with fixed probabilities (70% or 30%), whereas selection of the other first-stage option lead to each pair of second-stage options with the opposite probabilities (i.e., 30% and 70%). Then, selection of one of the second-stage options lead to reward or no-reward outcome. The probability of reward for each second-stage option was independently set according to Gaussian random walk with reflecting boundaries at 0.25 and 0.75. More specifically, the reward probabilities for the four second-stage options were independently changed at each trial by adding pseudo normal random numbers with mean 0 and SD 0.025, and reflected at 0.25 and 0.75, throughout the task consisting of 201 trials (Figure 10Ba).

We also simulated a variant of the task, where the probabilities of reward for the four second-stage options (two options for each of the two pairs) were set to specific values and the option-probability contingency was changed three times during the task consisting of 201 trials. Specifically, the probabilities were initially set to (0.1 and 0.5) for the first and second option of the first pair and (0.5 and 0.9) for the first and second option of the second pair, respectively, and changed to (0.9 and 0.5) and (0.5 and 0.1) at the 51th trial, (0.5 and 0.1) and (0.9 and 0.5) at the 101th trial, and (0.5 and

0.9) and (0.1 and 0.5) at the 151th trial (Figure 10Bb).

We further simulated another variant of the task, in which there were three, rather than two, first-stage options and three pairs of second-stage options (Figure 10D, left). Selection of each of the three first-stage options lead to one of three pairs of second-stage options with fixed probabilities: (60%, 20%, 20%), (20%, 60%, 20%), and (20%, 20%, 60%). The probabilities of reward for the six second-stage options (two options for each of the three pairs) were set to specific values and the option-probability contingency was changed two times during the task consisting of 150 trials. Specifically, the probabilities were initially set to (0.5, 0.9), (0.1, 0.5), and (0.1, 0.5) for the (first, second) option of the first, second, and third pair, and changed to (0.5, 0.1), (0.9, 0.5), and (0.5, 0.1) at the 51th trial, and (0.1, 0.5), (0.1, 0.5), and (0.5, 0.9) at the 101th trial (Figure 10D, right).

In all the variants of the two-stage tasks, at every trial, SARSA-type TD RPE for the first stage was calculated after the second-stage choice was determined:

$$\delta_k(1) = 0 + \gamma V(O_k(2)) - V(O_k(1)),$$

where $O_k(1)$ and $O_k(2)$ represent the chosen option for the first and second stage at the k-th trial, respectively. $V(O)$ represents the value of option O. γ is the time discount factor, which was assumed to be 1 (i.e., no temporal discounting) in the two-stage tasks. Then TD RPE for the second stage:

$$\delta_k(2) = R + 0 - V(O_k(2))$$

was calculated after reward (R = 1 or 0) was determined, where $V(O_k(2))$ reflected $\delta_k(1)$-based updates of the weights of the approximate value function (see below) in the cases with SR-based system(s). The IR-based system(s) learned the value of each first- and second-stage option (in total 2+4=6 options or 3+6=9 options) through RPE-based updates, with the initial values set to 0. The SR-based system(s) had the SR of the first- and second-stage options, and leaned their values through RPE-based updates of the weights of the approximate value function. The SR matrix (6×6 or 9×9) was initialized to the one under the random policy regarding the choice at the second stage, incorporating the presumed stage-transition probabilities. The SR matrix was then updated by using the prediction errors of SR features. More specifically, at every trial, the SR features for the option chosen at the first stage was updated according to the prediction errors after the stage transition occurred and the second-stage option was determined, with the learning rate set to 0.05. Choice at both stages was made in the soft-max manner with the degree of exploitation over exploration (inverse temperature) (β) set to 5.

Simulations and statistics

Simulations were conducted n = 100 times for the reward navigation tasks, or 1000 times for the two-stage tasks, for each condition by using MATLAB. Probabilities and pseudorandom numbers were implemented by "rand", "randn", and "randperm" functions of MATLAB. Standard deviation (SD) was calculated with normalization by n, and standard error of the mean (SEM) was approximately calculated as SD/$\sqrt{n}$. For the conditions shown in Figure 3C, 50 sets of 100 simulations (including the

one shown in Figure 3A,B) were conducted.

Code availability

Codes to generate the data presented in the figures and tables are available at: https://github.com/kenjimoritagithub/sr2

Acknowledgements

KM was supported by Grant-in-Aid for Scientific Research (No. 20H05049) of the Ministry of Education, Culture, Sports, Science and Technology in Japan (http://www.mext.go.jp/en/). YK was supported by Grant-in-Aid for Scientific Research (No. 17H06311, 20H03359) and by the Cooperative Study Program of National Institute for Physiological Sciences (Section of Electron Microscopy).

References

1. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004;306(5703):1940-3. doi: 1102941 [pii] 10.1126/science.1102941. PubMed PMID: 15528409.
2. 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. 2010;66(6):896-907. doi: S0896-6273(10)00379-X [pii] 10.1016/j.neuron.2010.05.011. PubMed PMID: 20620875.
3. Kravitz AV, Tye LD, Kreitzer AC. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat Neurosci. 2012;15(6):816-8. doi: nn.3100 [pii] 10.1038/nn.3100. PubMed PMID: 22544310.
4. Iino Y, Sawada T, Yamaguchi K, Tajiri M, Ishii S, Kasai H, et al. Dopamine D2 receptors in discrimination learning and spine enlargement. Nature. 2020;579(7800):555-60. Epub 20200318. doi: 10.1038/s41586-020-2115-1. PubMed PMID: 32214250.
5. Lee SJ, Lodder B, Chen Y, Patriarchi T, Tian L, Sabatini BL. Cell-type-specific asynchronous modulation of PKA by dopamine in learning. Nature. 2021;590(7846):451-6. Epub 20201223. doi: 10.1038/s41586-020-03050-5. PubMed PMID: 33361810; PubMed Central PMCID: PMCPMC7889726.
6. Groman SM, Keistler C, Keip AJ, Hammarlund E, DiLeone RJ, Pittenger C, et al. Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. Neuron. 2019;103(4):734-46.e3. Epub 20190625. doi: 10.1016/j.neuron.2019.05.042. PubMed PMID: 31253468; PubMed Central PMCID: PMCPMC6893860.
7. Collins AG, Frank MJ. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev. 2014;121(3):337-66. doi: 10.1037/a0037015. PubMed PMID: 25090423.
8. Möller M, Bogacz R. Learning the payoffs and costs of actions. PLoS Comput Biol. 2019;15(2):e1006285. Epub 2019/02/28. doi: 10.1371/journal.pcbi.1006285. PubMed PMID: 30818357; PubMed Central PMCID: PMCPMC6413954.
9. Mikhael JG, Bogacz R. Learning Reward Uncertainty in the Basal Ganglia. PLoS Comput Biol. 2016;12(9):e1005062. Epub 2016/09/02. doi: 10.1371/journal.pcbi.1005062. PubMed PMID: 27589489; PubMed Central PMCID: PMCPMC5010205.
10. Chen Y, Monaco S, Byrne P, Yan X, Henriques DY, Crawford JD. Allocentric versus egocentric representation of remembered reach targets in human cortex. J Neurosci. 2014;34(37):12515-26. doi: 10.1523/JNEUROSCI.1445-14.2014. PubMed PMID: 25209289; PubMed Central PMCID: PMCPMC6615499.
11. Town SM, Brimijoin WO, Bizley JK. Egocentric and allocentric representations in auditory cortex.

PLoS Biol. 2017;15(6):e2001878. Epub 20170615. doi: 10.1371/journal.pbio.2001878. PubMed PMID: 28617796; PubMed Central PMCID: PMCPMC5472254.

12. Wang C, Chen X, Knierim JJ. Egocentric and allocentric representations of space in the rodent brain. Curr Opin Neurobiol. 2020;60:12-20. Epub 20191130. doi: 10.1016/j.conb.2019.11.005. PubMed PMID: 31794917; PubMed Central PMCID: PMCPMC7080648.
13. Ballion B, Mallet N, Bézard E, Lanciego JL, Gonon F. Intratelencephalic corticostriatal neurons equally excite striatonigral and striatopallidal neurons and their discharge activity is selectively reduced in experimental parkinsonism. Eur J Neurosci. 2008;27(9):2313-21. doi: EJN6192 [pii] 10.1111/j.1460-9568.2008.06192.x. PubMed PMID: 18445222.
14. Kress GJ, Yamawaki N, Wokosin DL, Wickersham IR, Shepherd GM, Surmeier DJ. Convergent cortical innervation of striatal projection neurons. Nat Neurosci. 2013;16(6):665-7. doi: 10.1038/nn.3397. PubMed PMID: 23666180.
15. Lei W, Jiao Y, Del Mar N, Reiner A. Evidence for differential cortical input to direct pathway versus indirect pathway striatal projection neurons in rats. J Neurosci. 2004;24(38):8289-99. doi: 24/38/8289 [pii] 10.1523/JNEUROSCI.1990-04.2004. PubMed PMID: 15385612.
16. Reiner A, Hart NM, Lei W, Deng Y. Corticostriatal projection neurons - dichotomous types and dichotomous functions. Front Neuroanat. 2010;4:142. doi: 10.3389/fnana.2010.00142. PubMed PMID: 21088706; PubMed Central PMCID: PMCPMC2982718.
17. Morita K. Differential cortical activation of the striatal direct and indirect pathway cells: reconciling the anatomical and optogenetic results by using a computational method. J Neurophysiol. 2014;112(1):120-46. doi: 10.1152/jn.00625.2013. PubMed PMID: 24598515.
18. Wall NR, De La Parra M, Callaway EM, Kreitzer AC. Differential innervation of direct- and indirect-pathway striatal projection neurons. Neuron. 2013;79(2):347-60. doi: 10.1016/j.neuron.2013.05.014. PubMed PMID: 23810541; PubMed Central PMCID: PMCPMC3729794.
19. Sutton R, Barto A. Reinforcement Learning. Cambridge, MA: MIT Press; 1998.
20. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol. 2017;13(9):e1005768. Epub 2017/09/25. doi: 10.1371/journal.pcbi.1005768. PubMed PMID: 28945743; PubMed Central PMCID: PMCPMC5628940.
21. Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ. The successor representation in human reinforcement learning. Nat Hum Behav. 2017;1(9):680-92. Epub 2017/08/28. doi: 10.1038/s41562-017-0180-8. PubMed PMID: 31024137; PubMed Central PMCID: PMCPMC6941356.
22. Garvert MM, Dolan RJ, Behrens TE. A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. Elife. 2017;6:e17086. Epub 2017/04/27. doi: 10.7554/eLife.17086. PubMed PMID: 28448253; PubMed Central PMCID: PMCPMC5407855.
23. Stachenfeld KL, Botvinick MM, Gershman SJ. The hippocampus as a predictive map. Nat

Neurosci. 2017;20(11):1643-53. Epub 2017/10/02. doi: 10.1038/nn.4650. PubMed PMID: 28967910.

24. Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. Neural evidence for the successor representation in choice evaluation. bioRxiv. 2021;https://doi.org/10.1101/2021.08.29.458114. doi: https://doi.org/10.1101/2021.08.29.458114.
25. Dayan P. Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Computation. 1993;5(4):613-24.
26. Balleine BW, O'Doherty JP. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 2010;35(1):48-69. doi: 10.1038/npp.2009.131. PubMed PMID: 19776734; PubMed Central PMCID: PMCPMC3055420.
27. Dolan RJ, Dayan P. Goals and habits in the brain. Neuron. 2013;80(2):312-25. doi: 10.1016/j.neuron.2013.09.007. PubMed PMID: 24139036; PubMed Central PMCID: PMCPMC3807793.
28. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011;69(6):1204-15. doi: S0896-6273(11)00125-5 [pii] 10.1016/j.neuron.2011.02.027. PubMed PMID: 21435563; PubMed Central PMCID: PMCPMC3077926.
29. Groman SM, Massi B, Mathias SR, Curry DW, Lee D, Taylor JR. Neurochemical and Behavioral Dissections of Decision-Making in a Rodent Multistage Task. J Neurosci. 2019;39(2):295-306. Epub 20181109. doi: 10.1523/JNEUROSCI.2219-18.2018. PubMed PMID: 30413646; PubMed Central PMCID: PMCPMC6325257.
30. Kool W, Cushman FA, Gershman SJ. When Does Model-Based Control Pay Off? PLoS Comput Biol. 2016;12(8):e1005090. Epub 20160826. doi: 10.1371/journal.pcbi.1005090. PubMed PMID: 27564094; PubMed Central PMCID: PMCPMC5001643.
31. Lehnert L, Tellex S, Littman ML. Advantages and Limitations of using Successor Features for Transfer in Reinforcement Learning. arXiv. 2017:arXiv:1708.00102v1.
32. Barreto A, Dabney W, Munos R, Hunt JJ, Schaul T, van Hasselt H, et al. Successor Features for Transfer in Reinforcement Learning. arXiv:160605312. 2016.
33. Piray P, Daw ND. Linear reinforcement learning in planning, grid fields, and cognitive control. Nat Commun. 2021;12(1):4942. Epub 20210816. doi: 10.1038/s41467-021-25123-3. PubMed PMID: 34400622; PubMed Central PMCID: PMCPMC8368103.
34. Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci. 2012;15(9):1281-9. doi: 10.1038/nn.3188. PubMed PMID: 22902719; PubMed Central PMCID: PMCPMC3951287.
35. Kim HF, Amita H, Hikosaka O. Indirect Pathway of Caudal Basal Ganglia for Rejection of Valueless Visual Objects. Neuron. 2017;94(4):920-30.e3. doi: 10.1016/j.neuron.2017.04.033.

PubMed PMID: 28521141.
36. Nonomura S, Nishizawa K, Sakai Y, Kawaguchi Y, Kato S, Uchigashima M, et al. Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. Neuron. 2018. doi: https://doi.org/10.1016/j.neuron.2018.08.002.
37. Matamales M, McGovern AE, Mi JD, Mazzone SB, Balleine BW, Bertran-Gonzalez J. Local D2- to D1-neuron transmodulation updates goal-directed learning in the striatum. Science. 2020;367(6477):549-55. doi: 10.1126/science.aaz5751. PubMed PMID: 32001651.
38. Peak J, Chieng B, Hart G, Balleine BW. Striatal direct and indirect pathway neurons differentially control the encoding and updating of goal-directed learning. Elife. 2020;9. Epub 20201120. doi: 10.7554/eLife.58544. PubMed PMID: 33215609; PubMed Central PMCID: PMCPMC7707820.
39. Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science. 2008;321(5890):848-51. doi: 321/5890/848 [pii] 10.1126/science.1160575. PubMed PMID: 18687967; PubMed Central PMCID: PMCPMC2833421.
40. Yagishita S, Hayashi-Takagi A, Ellis-Davies GC, Urakubo H, Ishii S, Kasai H. A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science. 2014;345(6204):1616-20. doi: 10.1126/science.1255514. PubMed PMID: 25258080.
41. Cowan RL, Wilson CJ. Spontaneous firing patterns and axonal projections of single corticostriatal neurons in the rat medial agranular cortex. J Neurophysiol. 1994;71(1):17-32. PubMed PMID: 8158226.
42. Hart G, Bradfield LA, Balleine BW. Prefrontal Corticostriatal Disconnection Blocks the Acquisition of Goal-Directed Action. J Neurosci. 2018;38(5):1311-22. Epub 20180104. doi: 10.1523/JNEUROSCI.2850-17.2017. PubMed PMID: 29301872; PubMed Central PMCID: PMCPMC6596266.
43. Hart G, Bradfield LA, Fok SY, Chieng B, Balleine BW. The Bilateral Prefronto-striatal Pathway Is Necessary for Learning New Goal-Directed Actions. Curr Biol. 2018;28(14):2218-29.e7. Epub 20180628. doi: 10.1016/j.cub.2018.05.028. PubMed PMID: 30056856.
44. Morishima M, Morita K, Kubota Y, Kawaguchi Y. Highly differentiated projection-specific cortical subnetworks. J Neurosci. 2011;31(28):10380-91.
45. Morita K, Im S, Kawaguchi Y. Differential Striatal Axonal Arborizations of the Intratelencephalic and Pyramidal-Tract Neurons: Analysis of the Data in the MouseLight Database. Front Neural Circuits. 2019;13:71. Epub 20191115. doi: 10.3389/fncir.2019.00071. PubMed PMID: 31803027; PubMed Central PMCID: PMCPMC6872499.
46. Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, et al. Concurrent activation of striatal direct and indirect pathways during action initiation. Nature. 2013;494(7436):238-42. Epub 2013/01/23. doi: 10.1038/nature11846. PubMed PMID: 23354054; PubMed Central PMCID: PMCPMC4039389.
47. Averbeck B, O'Doherty JP. Reinforcement-learning in fronto-striatal circuits.

Neuropsychopharmacology. 2021:10.1038/s41386-021-01108-0. Epub 20210805. doi: 10.1038/s41386-021-01108-0. PubMed PMID: 34354249.

48. Morita K, Morishima M, Sakai K, Kawaguchi Y. Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. Trends Neurosci. 2012;35(8):457–67. doi: 10.1016/j.tins.2012.04.009.
49. Morita K, Morishima M, Sakai K, Kawaguchi Y. Dopaminergic control of motivation and reinforcement learning: a closed-circuit account for reward-oriented behavior. J Neurosci. 2013;33(20):8866-90. doi: 10.1523/JNEUROSCI.4614-12.2013. PubMed PMID: 23678129.
50. Watabe-Uchida M, Eshel N, Uchida N. Neural Circuitry of Reward Prediction Error. Annu Rev Neurosci. 2017;40:373-94. Epub 2017/04/24. doi: 10.1146/annurev-neuro-072116-031109. PubMed PMID: 28441114.
51. Gershman SJ, Moore CD, Todd MT, Norman KA, Sederberg PB. The successor representation and temporal context. Neural Comput. 2012;24(6):1553-68. Epub 2012/02/24. doi: 10.1162/NECO_a_00282. PubMed PMID: 22364500.
52. Gardner MPH, Schoenbaum G, Gershman SJ. Rethinking dopamine as generalized prediction error. Proc Biol Sci. 2018;285(1891):20181645. Epub 2018/11/21. doi: 10.1098/rspb.2018.1645. PubMed PMID: 30464063; PubMed Central PMCID: PMCPMC6253385.

Tables

Table 1

Performance of the models consisting of two systems in a broader parameter space. The left, middle, and right part shows the results for the model consisting of SR- and IR-based systems, two SR-based systems, and two IR-based systems, respectively. Each sub-table shows the sets of learning rate parameters that gave top ten mean performance for each set of time discount factor (γ) and inverse temperature (β) (shown in the left) in each of the three models. For the model consisting of SR- and IR-based systems, cases with a combination of appetitive ($\alpha_+/\alpha_- > 1$) SR-based system and aversive ($\alpha_+/\alpha_- < 1$) IR-based system are shown in red; notably, even in all the other cases shown in this table except for a case shown in light blue, the α_+/α_- ratio was higher in the SR-based system than in the IR-based system.

Table 2

Results of additional simulations for the model consisting of two SR-based systems. Each sub-table shows the sets of learning rate parameters that gave top ten mean performance for each set of time discount factor (γ) and inverse temperature (β) shown in the left.

Figure legends

Figure 1

The simulated reward navigation task. **(A)** The 5 × 5 grid space, where the agent moved around. The agent started from the fixed start state, (1, 1), and moved to one of the neighboring states (two states at the corners, three states at the edges, and four states elsewhere) at each time step. There were nine reward candidate states, where reward could potentially be placed, namely, (1, 5), (2, 5), (3, 5), (4, 5), (5, 1), (5, 2), (5, 3), (5, 4), and (5, 5) (indicated by the gray color). **(B)** Epochs in the task. During the initial 500 time steps, there was no reward, and this was called the no-reward epoch. During the next 500 time steps, one of the nine reward candidate states was specified as a special candidate state, whereas the remaining eight reward candidate states were regarded as normal candidate states. There were in total nine rewarded epochs (500 time steps for each), and each one of the nine reward candidate states became the special candidate state in one of the nine epochs; the order was determined by pseudorandom permutation in each single simulation. **(C)** In the rewarded epochs, if the agent reached the rewarded state and obtained the reward, the agent was immediately carried back to the start state, and a new reward was introduced into a state, which was the special reward candidate state with 60% probability and one of the eight normal candidate states with 5% probability for each.

Figure 2

The reinforcement learning model with two systems. **(A)** Model architecture. The model consists of two learning systems, system 1 and system 2, which may use different ways of state representations. Each system has its own system-specific value of each state, and their mean (average) becomes the integrated state value. The agent selects an action to move to a neighboring state depending on their integrated values in a soft-max manner. The integrated values are also used for calculating the temporal-difference reward prediction error (TD-RPE). The TD-RPE was then used for updating the system-specific state values in each of the two systems. The leaning rates of the updates of the system-specific state values can differ between the two systems and also depending on whether the TD-RPE is positive (non-negative) or negative. **(B)** Combinations of the ways of state representation in the two systems. Combination of the successor representation (SR) in system 1 and the individual representation (IR) in system 2 was examined in Figures 3-5, 7, 9 and 10. Cases with only the SR or IR in both systems were examined in Figures 6 and 8.

Figure 3

Performance of the model consisting of a system using the SR and another system using the IR, and its dependence on the ratios of the learning rates from positive and negative TD-RPEs in each system. **(A)** Mean performance over $n = 100$ simulations for each condition. The axis rising to the left indicates the ratio of positive- / negative-error-based learning rates (denoted as α_+/α_-) in the system using the

SR, while the axis rising to the right indicates the same ratio in the system using the IR. The sum of the learning rates from positive and negative TD-RPEs ($\alpha_+ + \alpha_-$) in each system was constant at 1 in any conditions shown in this figure. The inverse temperature β was 5, and the time discount factor γ was 0.7. The vertical line corresponds to conditions where the α_+/α_- ratio is equal for both systems (bottom: negative error-based learning dominates; top: positive error-based learning dominates). The left side of the vertical line corresponds to conditions where the α_+/α_- ratio is larger in the SR-based system than in the IR-based system, whereas the opposite is the case for the right side of the vertical line. The color in each square pixel indicates the mean total obtained rewards in the task, averaged across 100 simulations, for each condition (i.e., set of $\alpha_{SR+}/\alpha_{SR-}$ and $\alpha_{IR+}/\alpha_{IR-}$ at the center of the square pixel), in reference to the rightmost color bar. The white cross indicates the set of α_+/α_- ratios that gave the best performance ($\alpha_+/\alpha_- = 4$ and 0.25 for the SR- and IR-based systems, respectively) among those examined in this figure. Note that the minimum value of the color bar is not 0. Also, the maximum and minimum values of the color bar do not match the highest and lowest performances in this figure; rather, they were set so that the highest and lowest performances in all the simulations of the same task with different parameters and/or model architecture (i.e., not only SR+IR but also SR+SR and IR+IR) shown in this figure and Figures 4 and 6 can be covered. **(B)** The black solid line, gray thin error-bars, and black thick error-bars respectively show the mean, standard deviation (SD, normalized by n; same hereafter), and standard error of the mean (SEM, approximated by SD/$\sqrt{n}$; same hereafter) of the performance over $n = 100$ simulations for the conditions where α_+/α_- for SR system times α_+/α_- for IR system was equal to 1 (i.e., the conditions on the horizontal diagonal in (A)). **(C)** Frequency (number of times) that each condition gave the best mean performance over 100 simulations when 100 simulations were executed 50 times (including the one shown in (A,B)) for each condition.

Figure 4

Performance of the model consisting of an SR-based system and an IR-based system with various sets of parameters. The sum of the learning rates from positive and negative TD-RPEs ($\alpha_+ + \alpha_-$) was varied over 0.75, 1, and 1.25. The inverse temperature β was varied over 5 and 10. The time discount factor γ was varied over 0.7 and 0.8. The ratio of positive- / negative-error-based learning rates (α_+/α_-) was varied over 0.2, 0.25, 1/3, 0.5, 1, 2, 3, 4, and 5 for the cases with $\alpha_+ + \alpha_- = 0.75$ or 1, but 0.2 and 5 were omitted for the cases with $\alpha_+ + \alpha_- = 1.25$ in order to avoid learning rate larger than 1. The color-performance correspondence is the same as in Figure 3. The white cross in each panel indicates the set of α_+/α_- ratios that gave the best performance among those examined in that condition, and the gray number near the top of each panel indicates the best performance (total rewards). The panel with $\alpha_+ + \alpha_- = 1$, $\beta = 5$, and $\gamma = 0.7$ shows the same results as shown in Figure 3.

Figure 5

Learning profiles of the model consisting of an SR-based system and an IR-based system with different combinations of the ratios of positive- / negative-error-based learning rates in each system. Three cases

included in Figure 3 (where the sum of the learning rates from positive and negative TD-RPEs ($\alpha_+ + \alpha_-$) in each system was 1, the inverse temperature β was 5, and the time discount factor γ was 0.7) were analyzed. **(A)** Mean learning curves. The curves indicate the mean time (number of time steps) used for obtaining the first, second, third, .. reward placed in each rewarded epoch (horizontal axis), averaged across simulations and also across eight (from the second to the ninth) rewarded epochs. Only the cases in which reward was obtained in not smaller than a quarter of (25 out of total 100) simulations in all of the eight epochs were plotted. The brown, red, and blue-green curves correspond to the conditions with (α_+/α_- for SR-system, α_+/α_- for IR-system) = (4, 0.25), (1, 1), and (0.25, 4), respectively (the colors match those in the corresponding pixels in Figure 3). The error-bars indicate ±SD across the eight epochs (after taking the averages across simulations). **(B)** Examples of the system-specific state values and the integrated state values. Panels in (a), (b), and (c) correspond to the conditions with (α_+/α_- for SR-system, α_+/α_- for IR-system) = (4, 0.25), (1, 1), and (0.25, 4), respectively. The left panels show the system-specific state values (red line: SR-based system, blue line: IR-based system) and the integrated state values (black solid line). The horizontal axis indicates 25 states (1-5 correspond to (1,1)-(5,1) in Figure 1/5B-right; 6-10 correspond to (1,2)-(5,2), and so on). Values just after the agent obtained the last reward in the last (9th) rewarded epoch (indicated by the white crosses in the right panels) in single simulations, in which that reward was placed at the "special candidate state (with high probability of becoming a rewarded state: see the Methods)" in that epoch, were shown. Note the differences in the vertical scales between (a) or (c) and (b). The middle panels show only the integrated state values. The vertical axes in (a) and (c) were enlarged as compared to the left panels. The right panels also show the integrated state values as a heat map on the 5 × 5 grid space.

Figure 6

Performance of the model consisting of two systems, both of which employed the same way of state representation. Results in the case where both systems employed only the SR **(A)** or only the IR **(B)** are shown. Configurations are the same as those in Figure 4 (white cross: the best-performance set of α_+/α_- ratios; gray number: the best performance (total rewards)). If the α_+/α_- ratio was equal between both systems, the two systems behaved in exactly the same manner, and thus such conditions (on the vertical lines) were equivalent to having only a single system. Also, only the left (A) or right (B) side is shown, because "$\alpha_{1+} / \alpha_{1-} = 0.2$ and $\alpha_{2+} / \alpha_{2-} = 3$", for example, are equivalent to "$\alpha_{1+} / \alpha_{1-} = 3$ and $\alpha_{2+} / \alpha_{2-} = 0.2$" given that both systems employed the same representation. The left (A)-and right-(B) placement was made so as to facilitate visual comparisons with Figure 3. **(C)** Comparison of the maximum performances (total rewards) of the cases where one system employed the SR while the other employed the IR (purple), both systems employed the SR (red), and both systems employed the IR (blue) for each set of the time discount factor (γ), inverse temperature (β), and the sum of learning rates for positive and negative TD-RPEs ($\alpha_+ + \alpha_-$). The bar lengths indicate the mean performance over 100 simulations for the condition that gave the best mean performance, and the black thick and gray thin error-bars indicate ±SEM and ±SD for that condition, respectively.

Figure 7

Performance of the model consisting of SR-based and IR-based systems in a broader parameter space. Each row of panels shows the mean performance for each set of time discount factor (γ) and inverse temperature (β) (shown in the left), varying the learning rate for the update of SR features ($\alpha_{\mathrm{SRfeature}}$, shown in the top), projected onto the plane consisting of the ratios of the learning rates from positive and negative TD-RPEs in the two systems ($\alpha_{\mathrm{SR+}}/\alpha_{\mathrm{SR-}}$ and $\alpha_{\mathrm{IR+}}/\alpha_{\mathrm{IR-}}$). There were 25 cases with $\alpha_{\mathrm{SR+}}/\alpha_{\mathrm{SR-}} = \alpha_{\mathrm{IR+}}/\alpha_{\mathrm{IR-}} = 1$, which were difficult to draw on the single point (1, 1) in this projected plane and thus omitted. In the cases with $\alpha_{\mathrm{SR+}}/\alpha_{\mathrm{SR-}} = 1$ or $\alpha_{\mathrm{IR+}}/\alpha_{\mathrm{IR-}} = 1$ but not $\alpha_{\mathrm{SR+}}/\alpha_{\mathrm{SR-}} = \alpha_{\mathrm{IR+}}/\alpha_{\mathrm{IR-}} = 1$, there were five cases having the same set of ($\alpha_{\mathrm{SR+}}/\alpha_{\mathrm{SR-}}$, $\alpha_{\mathrm{IR+}}/\alpha_{\mathrm{IR-}}$) (because $\alpha_+/\alpha_- = 1$ corresponds to five cases with $\alpha_+ = \alpha_- = 0.2$, 0.35, 0.5, 0.65, or 0.8), and these five cases were drawn by concentric circles with larger radius corresponding to larger learning rate. In the cases other than the special cases mentioned just above, each case was drawn by a cross. The color of the symbols (cross or circle) indicates the mean performance, in reference to the color bar in the top right.

Figure 8

Performance of the model consisting of two SR-based systems or two IR-based systems in a broader parameter space. Each row of panels shows the mean performance for each set of time discount factor (γ) and inverse temperature (β) (shown in the left), varying the learning rate for the update of SR features ($\alpha_{\mathrm{SRfeature}}$, shown in the top) for the cases with two SR-based systems (five panels from the left), projected onto the plane consisting of the ratios of the learning rates from positive and negative TD-RPEs in the two systems (α_{1+}/α_{1-} and α_{2+}/α_{2-}). The cases with $\alpha_{1+}/\alpha_{1-} = \alpha_{2+}/\alpha_{2-} = 1$ were not drawn, and the cases with $\alpha_{1+}/\alpha_{1-} = 1$ or $\alpha_{2+}/\alpha_{2-} = 1$ but not $\alpha_{1+}/\alpha_{1-} = \alpha_{2+}/\alpha_{2-} = 1$ were drawn by concentric circles, and all the other cases were drawn by crosses, in a similar manner to Figure 7. The color of the symbols (cross or circle) indicates the mean performance, in reference to the color bar in the top right. Only the left or right side is shown for the cases with two SR- or IR-based systems, respectively, for the same reason as in Figure 6.

Figure 9

Performance of the model consisting of SR-based and IR-based system when task properties were changed. The model with the original set of parameters ($\alpha_{\mathrm{SR+}} + \alpha_{\mathrm{SR-}} = \alpha_{\mathrm{IR+}} + \alpha_{\mathrm{IR-}} = 1$, $\beta = 5$, and $\gamma = 0.7$) was used. The ranges of the color bars in this figure correspond to the ranges between the lowest and highest performances (i.e., the minimum and maximum total rewards) in the individual panels, and they are different from the color bars in the previous figures. **(A)** Performance (total rewards) for the cases where the probability of reward placement at the special candidate state was varied from 70% (leftmost panel), 80%, 90%, or 100% (rightmost). **(B)** Performance for the cases where reward location (state) was reset at every 500 (leftmost panel), 250, 100, or 50 (rightmost) time steps in the rewarded epochs; where reward was located was determined according to the original stochastic rule, i.e., reward

was placed at the special candidate state for the epoch with 60% and placed at one of the other (normal) candidate states with 5% each. **(C)** Performance for the case where there was only a single rewarded epoch with 4500 time steps following the initial 500 time steps no-reward epoch. **(D)** Performance for the case where stochasticity of reward placement within each epoch was removed and the duration of each rewarded epoch was shortened from 500 time steps to 100 time steps while number of rewarded epochs was increased from 9 to 45. **(E)** Performance for the case where special candidate state was abolished and reward was placed at each of the nine candidate states with equal probability (1/9 = 11.11..%). **(F)** Performance for the case where rewarded state was determined by a fixed order, namely, (5, 1), (5, 5), (1, 5), and again (5, 1), and these were repeated throughout the task.

Figure 10

Performance of the two-system models in the two-stage tasks. **(A)** Schematic diagram of the two-stage task. Selection of one of the two first-stage options lead to one of the two pairs of second-stage options with fixed probabilities. **(B)** Reward probabilities for the four second-stage options. **(a)** In the original task, the probability of for each option was independently set according to Gaussian random walk with reflecting boundaries at 0.25 and 0.75. An example is shown. **(b)** In a variant of the task, the probabilities for the four options were set to specific values, which changed three times in the task. **(C)** Mean performance (mean total rewards over n = 1000 simulations) of the model consisting of an SR-based system and an IR-based system, with the ratio of positive- / negative-error-based learning rates (α_+/α_-) for each system was varied under $\alpha_+ + \alpha_- = 1$, $\beta = 5$, and $\gamma = 1$. Panels (a) and (b) show the results for the tasks with the reward probabilities shown in (B)(a) and (B)(b), respectively. **(D)** *Left*: Schematic diagram of a variant of the two-stage task, in which there were three first-stage options and three pairs of second-stage options. *Right*: Reward probabilities for the six second-stage options, which were set to specific values and changed two times in the task. **(E-G)** *Top panels*: Mean performance of the model consisting of SR- and IR-based systems (E), two SR-based systems (F), or two IR-based systems (G), in the task variant with three first-stage options and three pairs of second-stage options. *Bottom graphs*: Mean (black solid line), SEM (black thick error-bars; though hardly visible), and SD (gray thin error-bars) of the performance over n = 1000 simulations for the conditions where α_+/α_- for SR system times α_+/α_- for IR system was equal to 1 (i.e., the conditions on the horizontal diagonal in the top panels).

Figure 11

Possible implementation of the SR-based and IR-based learning systems in the cortico-BG circuits. SR is used in the sensory and limbic cortices and/or intratelencephalic (IT) neurons, whereas IR is used in the primary motor cortex and/or pyramidal-tract (PT) neurons. Sensory and limbic cortices and/or IT neurons preferentially project to the D1/direct pathway, whereas primary motor cortex and/or PT neurons preferentially target the D2/indirect pathway. The D1/direct pathway learns mainly from positive TD-RPEs, whereas the D2/indirect pathway learns mainly from negative TD-RPEs. Overall,

the circuits implement the combination of SR-based system learning mainly from positive TD-RPEs and IR-based system learning mainly from negative TD-RPEs, which achieved the best performance in our simulated task.

(End)

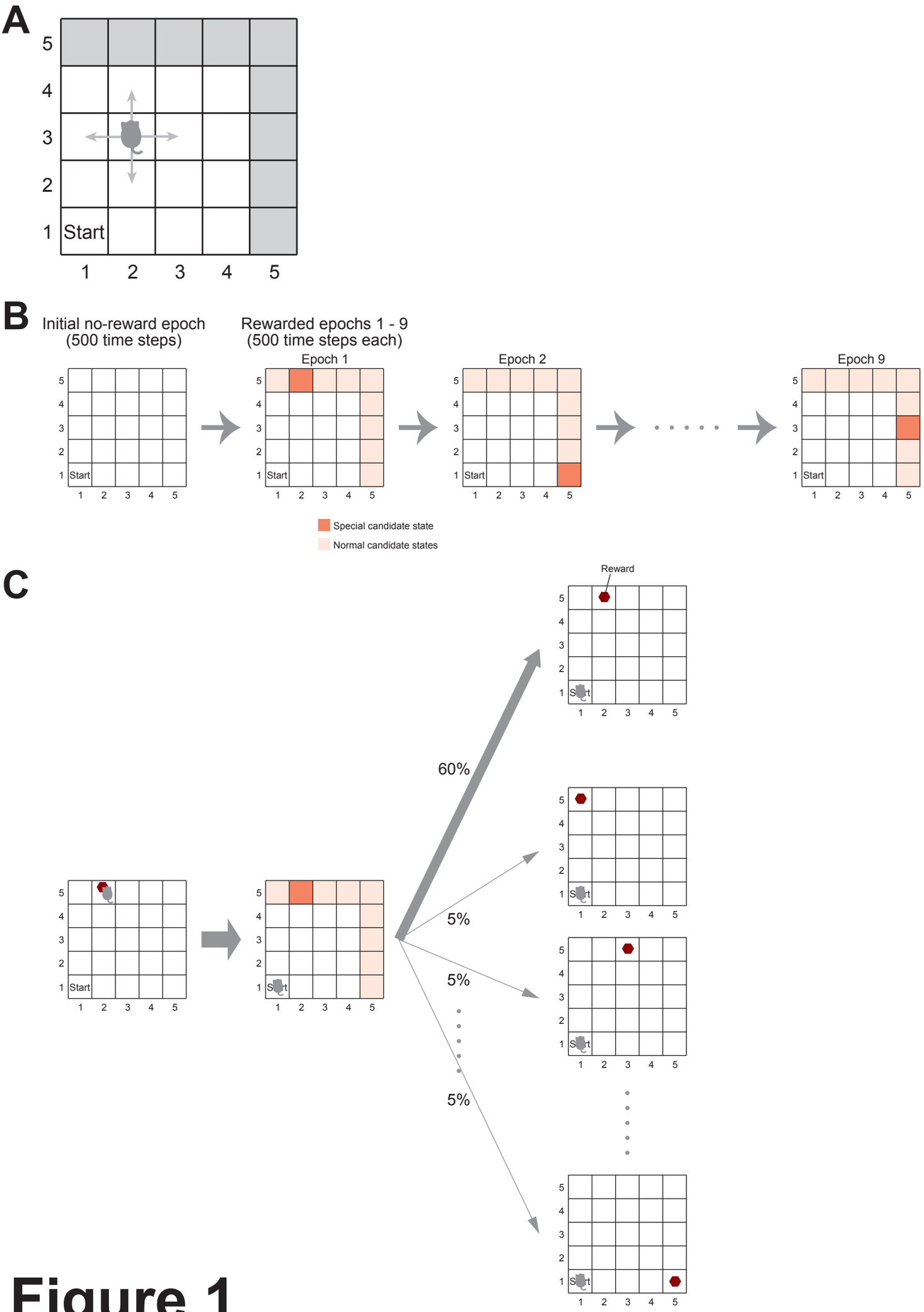

A
Start
B
Initial no-reward epoch
(500 time steps)
Rewarded epochs 1 - 9
(500 time steps each)
Epoch 1
Epoch 2
Epoch 9
Special candidate state
Normal candidate states
C
Reward
60%
5%
5%
5%

Figure 1

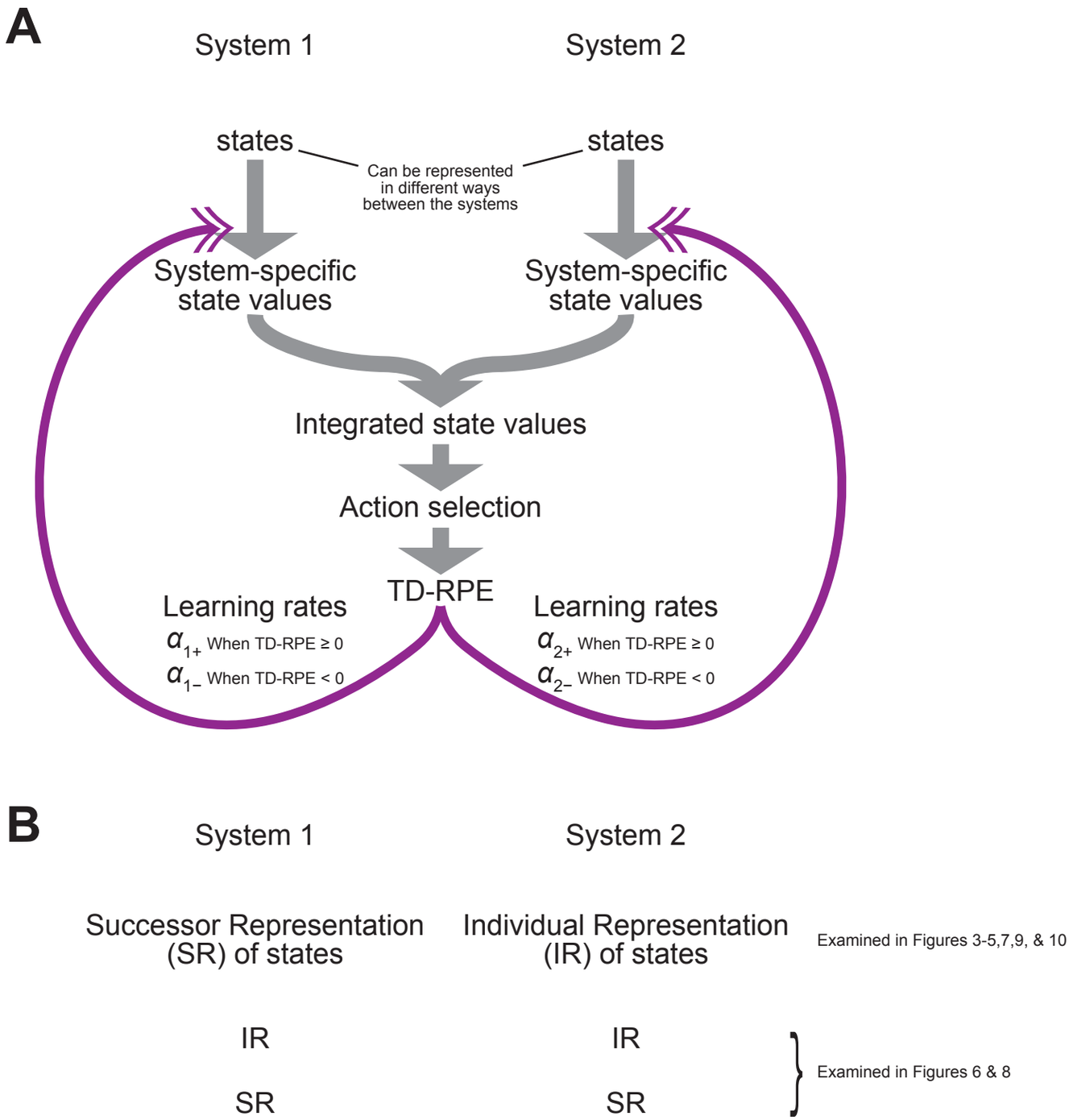

A
System 1
System 2
states
states
Can be represented in different ways between the systems
System-specific state values
System-specific state values
Integrated state values
Action selection
TD-RPE
Learning rates
α_{1+} When TD-RPE ≥ 0
α_{1-} When TD-RPE < 0
Learning rates
α_{2+} When TD-RPE ≥ 0
α_{2-} When TD-RPE < 0
B
System 1
System 2
Successor Representation (SR) of states
Individual Representation (IR) of states
Examined in Figures 3-5,7,9, & 10
IR
SR
IR
SR
Examined in Figures 6 & 8

Figure 2

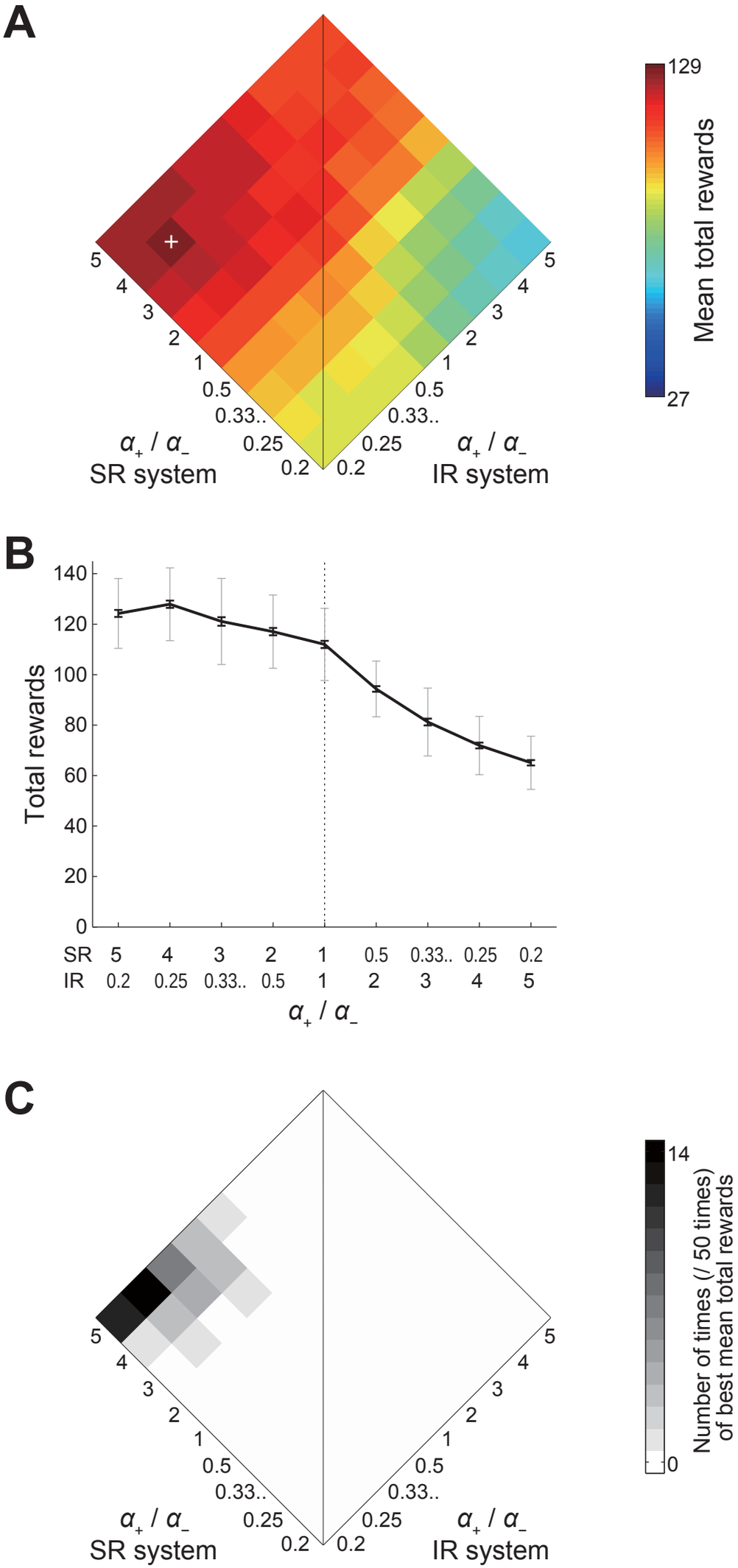

A
129
27
Mean total rewards
5
4
3
2
1
0.5
0.33..
0.25
0.2
α+ / α-
SR system
IR system
B
Total rewards
140
120
100
80
60
40
20
0
SR 5 4 3 2 1 0.5 0.33.. 0.25 0.2
IR 0.2 0.25 0.33.. 0.5 1 2 3 4 5
α+ / α-
C
14
0
Number of times (/ 50 times) of best mean total rewards

Figure 3

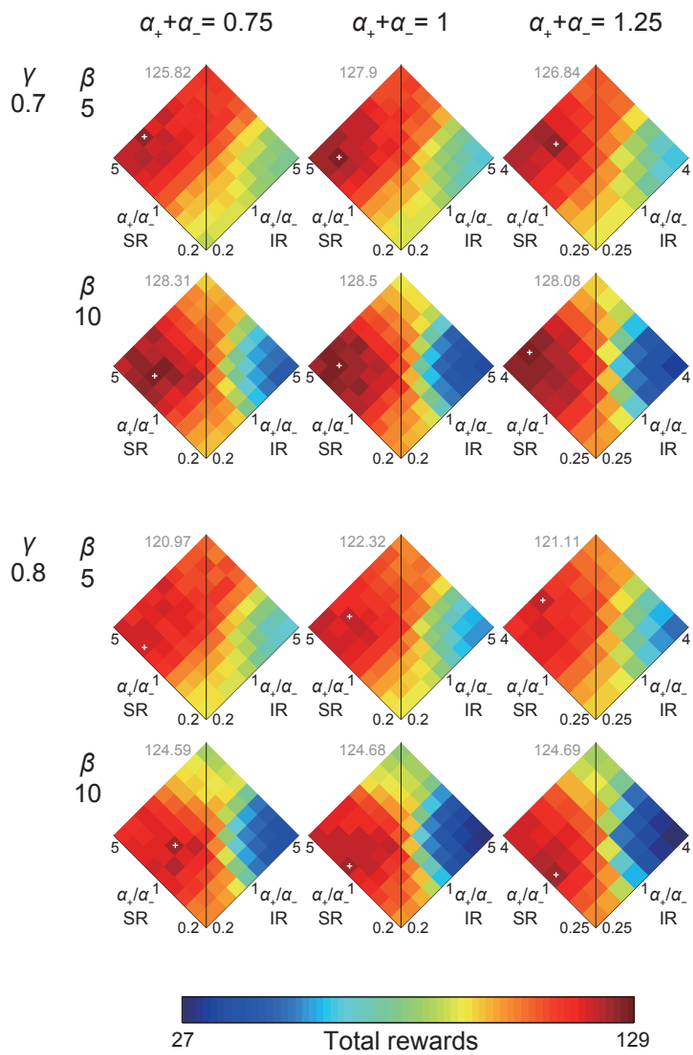

α+α- = 0.75
α+α- = 1
α+α- = 1.25
γ 0.7
β 5
β 10
γ 0.8
β 5
β 10
125.82
127.9
126.84
128.31
128.5
128.08
120.97
122.32
121.11
124.59
124.68
124.69
SR
IR
27
Total rewards
129

Figure 4

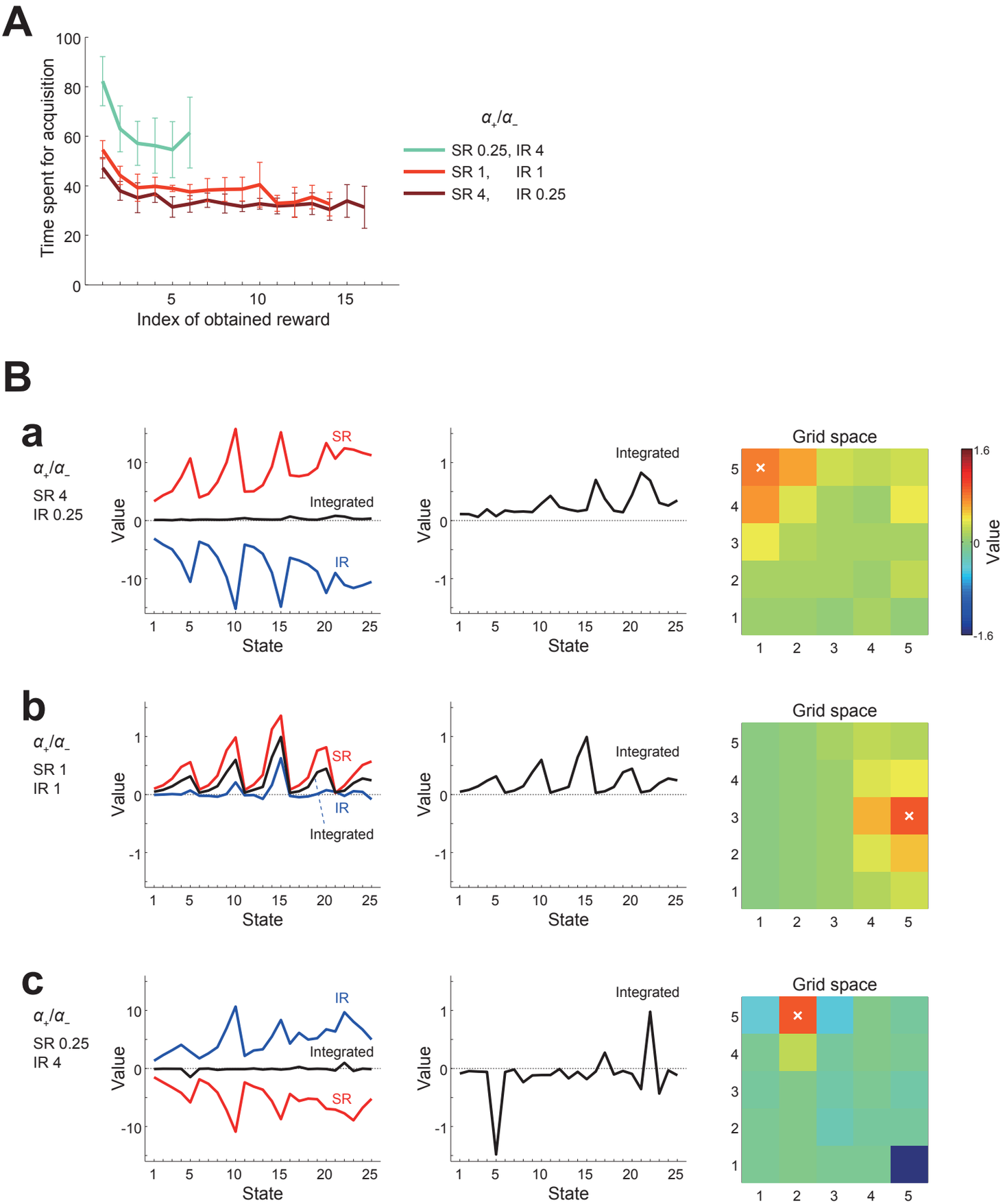
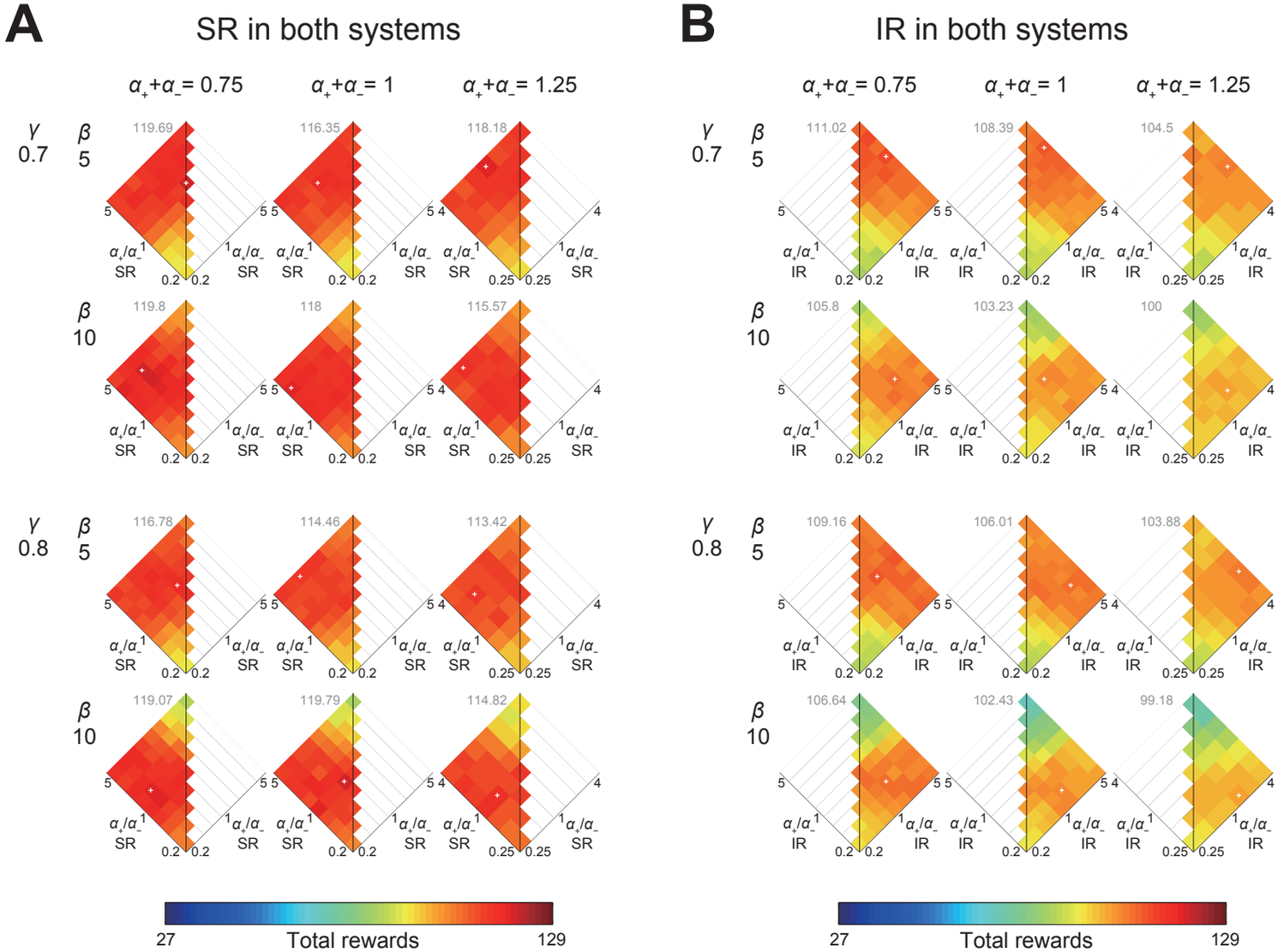

A
Time spent for acquisition
Index of obtained reward
α_+/α_-
SR 0.25, IR 4
SR 1, IR 1
SR 4, IR 0.25
B
a
α_+/α_-
SR 4
IR 0.25
SR
Integrated
IR
Value
State
Grid space
b
α_+/α_-
SR 1
IR 1
c
α_+/α_-
SR 0.25
IR 4

Figure 5


A
SR in both systems
B
IR in both systems
α+α- = 0.75
α+α- = 1
α+α- = 1.25
γ 0.7
β 5
β 10
γ 0.8
119.69
116.35
118.18
119.8
118
115.57
116.78
114.46
113.42
119.07
119.79
114.82
111.02
108.39
104.5
105.8
103.23
100
109.16
106.01
103.88
106.64
102.43
99.18
α+/α- SR
α+/α- IR
27
Total rewards
129

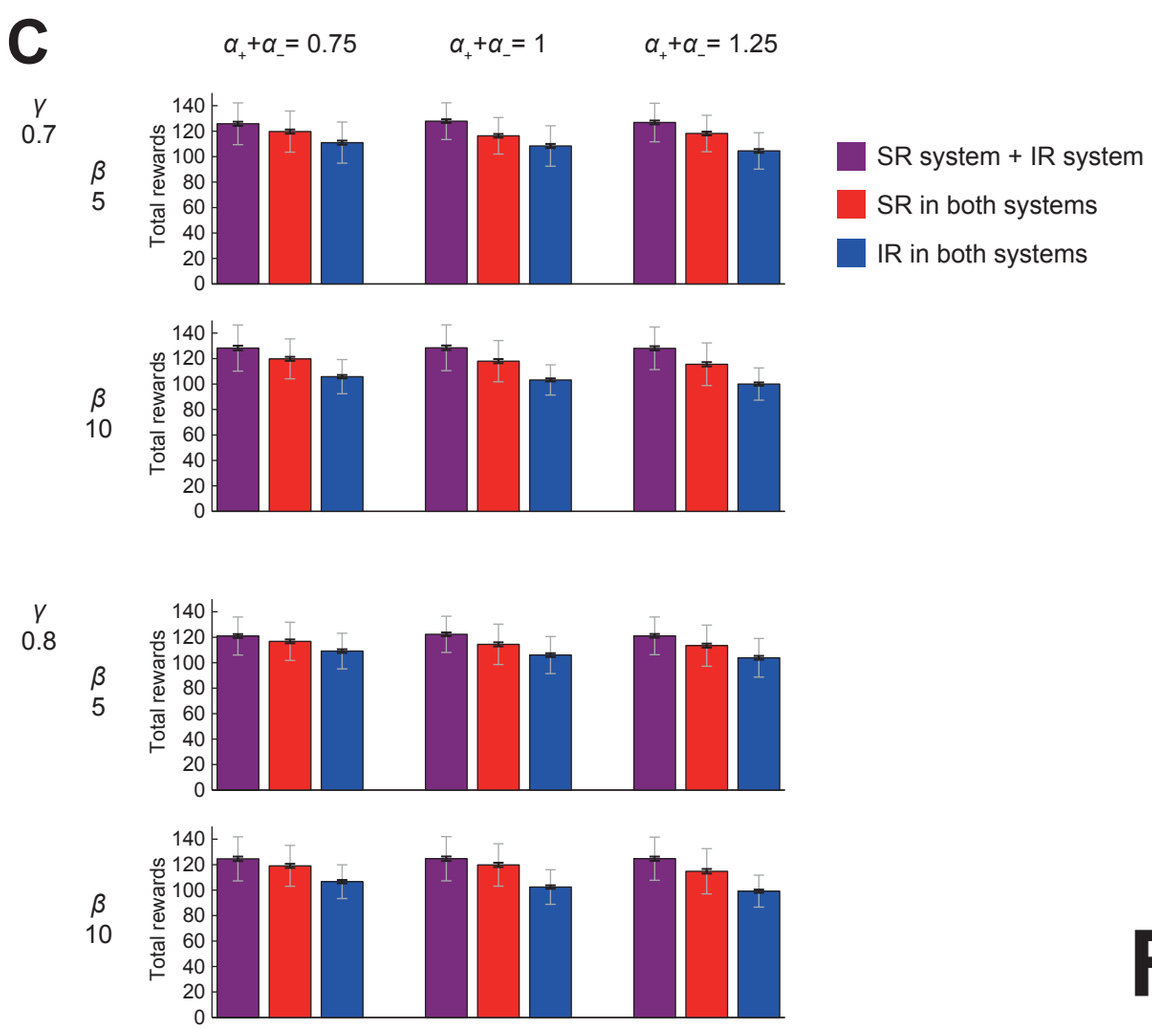

C
α+α- = 0.75
α+α- = 1
α+α- = 1.25
γ 0.7
γ 0.8
β 5
β 10
Total rewards
SR system + IR system
SR in both systems
IR in both systems

Figure 6

Figure 7

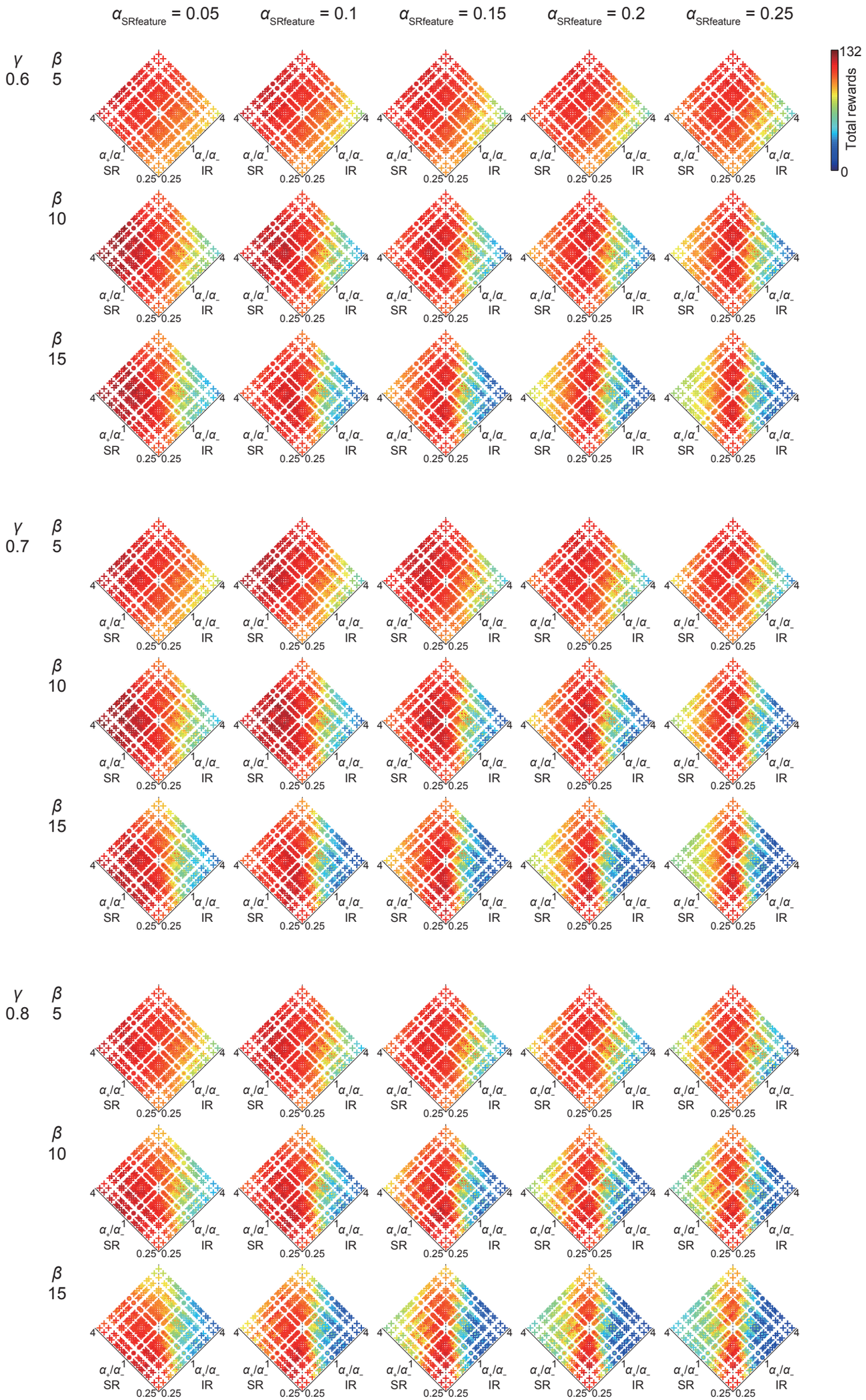
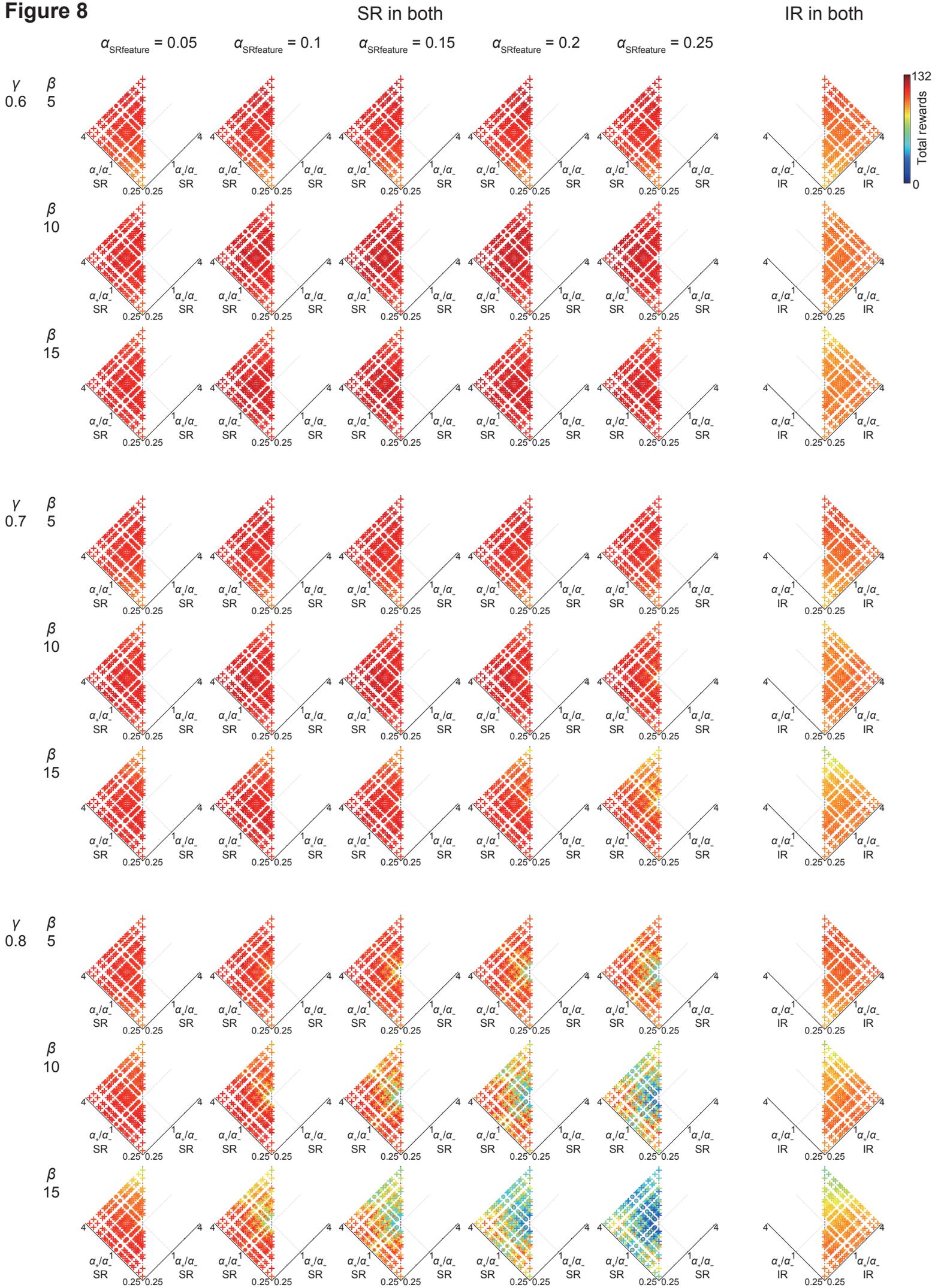

$\alpha_{SRfeature}$ = 0.05
$\alpha_{SRfeature}$ = 0.1
$\alpha_{SRfeature}$ = 0.15
$\alpha_{SRfeature}$ = 0.2
$\alpha_{SRfeature}$ = 0.25
γ 0.6
γ 0.7
γ 0.8
β 5
β 10
β 15
α_+/α_- SR
α_+/α_- IR
4
1
0.25
132
0
Total rewards

Figure 8


SR in both
IR in both
$\alpha_{SRfeature}$ = 0.05
$\alpha_{SRfeature}$ = 0.1
$\alpha_{SRfeature}$ = 0.15
$\alpha_{SRfeature}$ = 0.2
$\alpha_{SRfeature}$ = 0.25
γ 0.6
γ 0.7
γ 0.8
β 5
β 10
β 15
α_+/α_- SR
α_+/α_- IR
4
1
0.25
Total rewards
132
0

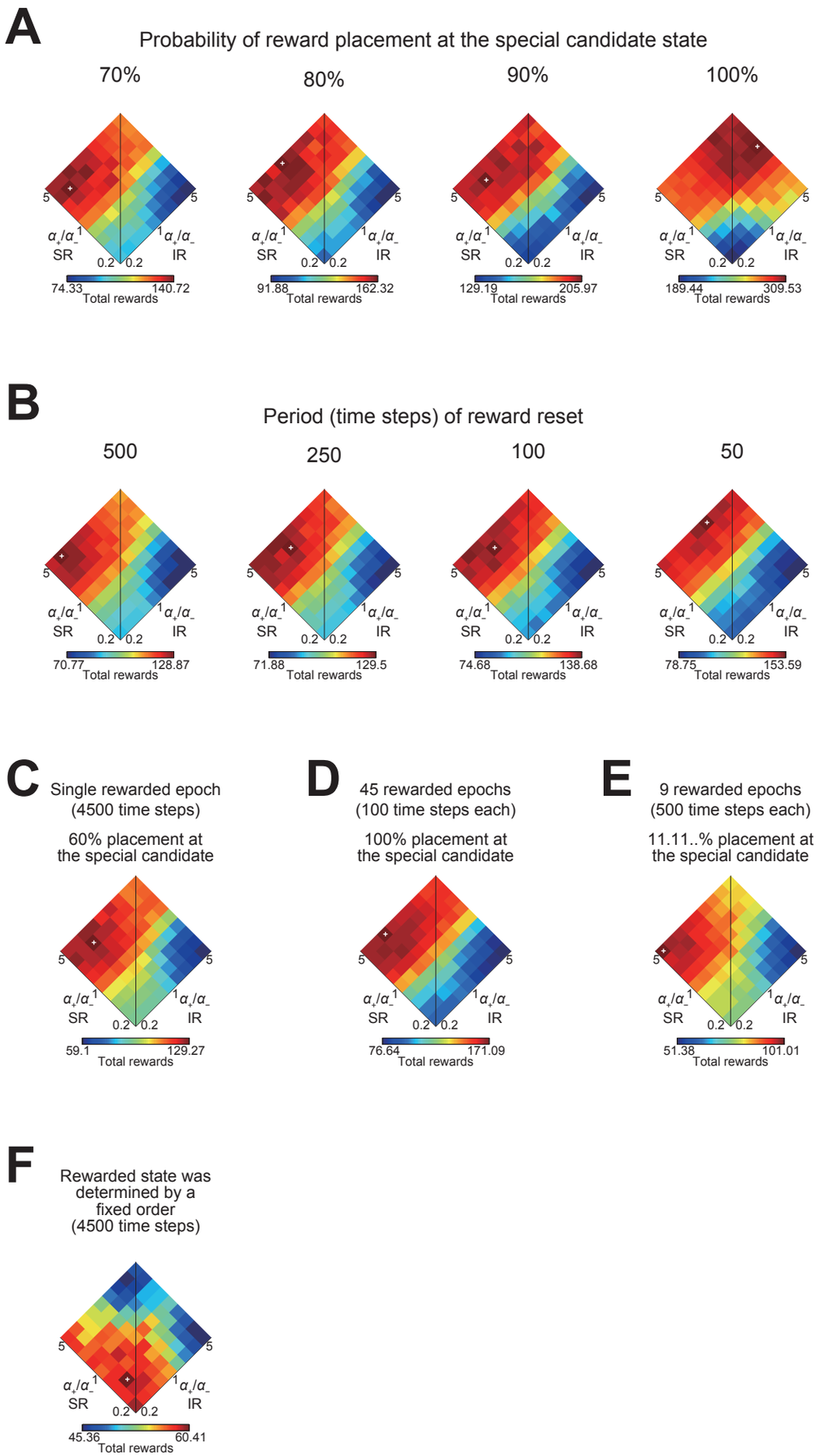
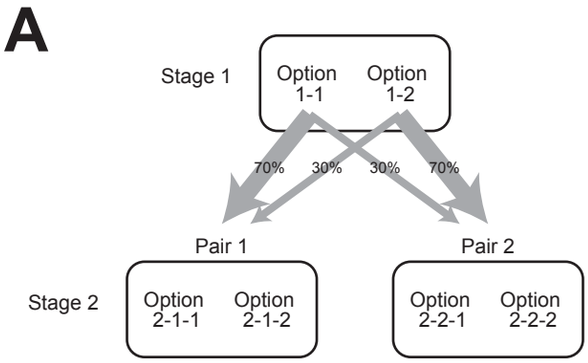
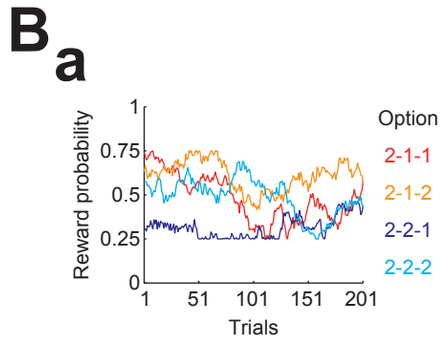

A
Probability of reward placement at the special candidate state
70%
80%
90%
100%
74.33
140.72
91.88
162.32
129.19
205.97
189.44
309.53
Total rewards
B
Period (time steps) of reward reset
500
250
100
50
70.77
128.87
71.88
129.5
74.68
138.68
78.75
153.59
Total rewards
C
Single rewarded epoch
(4500 time steps)
60% placement at
the special candidate
59.1
129.27
Total rewards
D
45 rewarded epochs
(100 time steps each)
100% placement at
the special candidate
76.64
171.09
Total rewards
E
9 rewarded epochs
(500 time steps each)
11.11..% placement at
the special candidate
51.38
101.01
Total rewards
F
Rewarded state was
determined by a
fixed order
(4500 time steps)
45.36
60.41
Total rewards
SR
IR
0.2
5

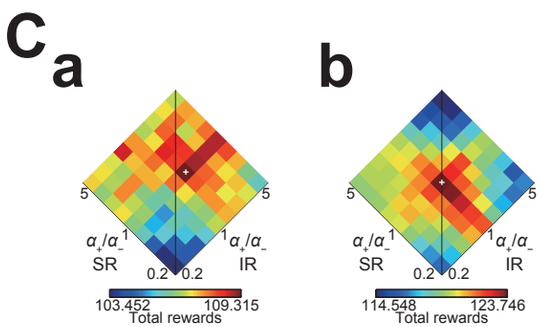
Figure 9

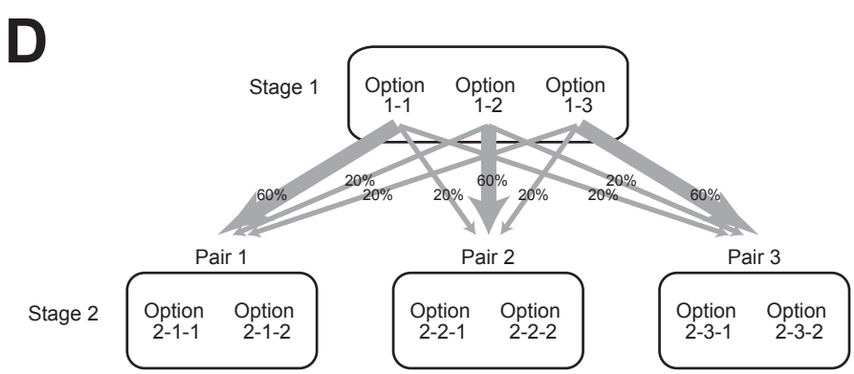




b

Option	Reward probability 1-	51-	101-	151- Trials
2-1-1	0.1	0.9	0.5	0.5
2-1-2	0.5	0.5	0.1	0.9
2-2-1	0.5	0.5	0.9	0.1
2-2-2	0.9	0.1	0.5	0.5





Option	Reward probability 1-	51-	101- Trials
2-1-1	0.5	0.5	0.1
2-1-2	0.9	0.1	0.5
2-2-1	0.1	0.9	0.1
2-2-2	0.5	0.5	0.5
2-3-1	0.1	0.5	0.9
2-3-2	0.5	0.1	0.5

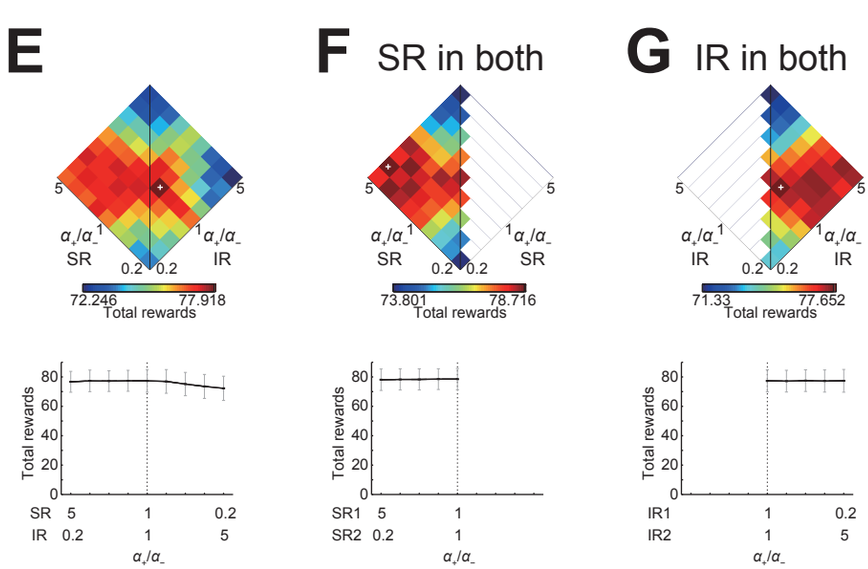


Figure 10

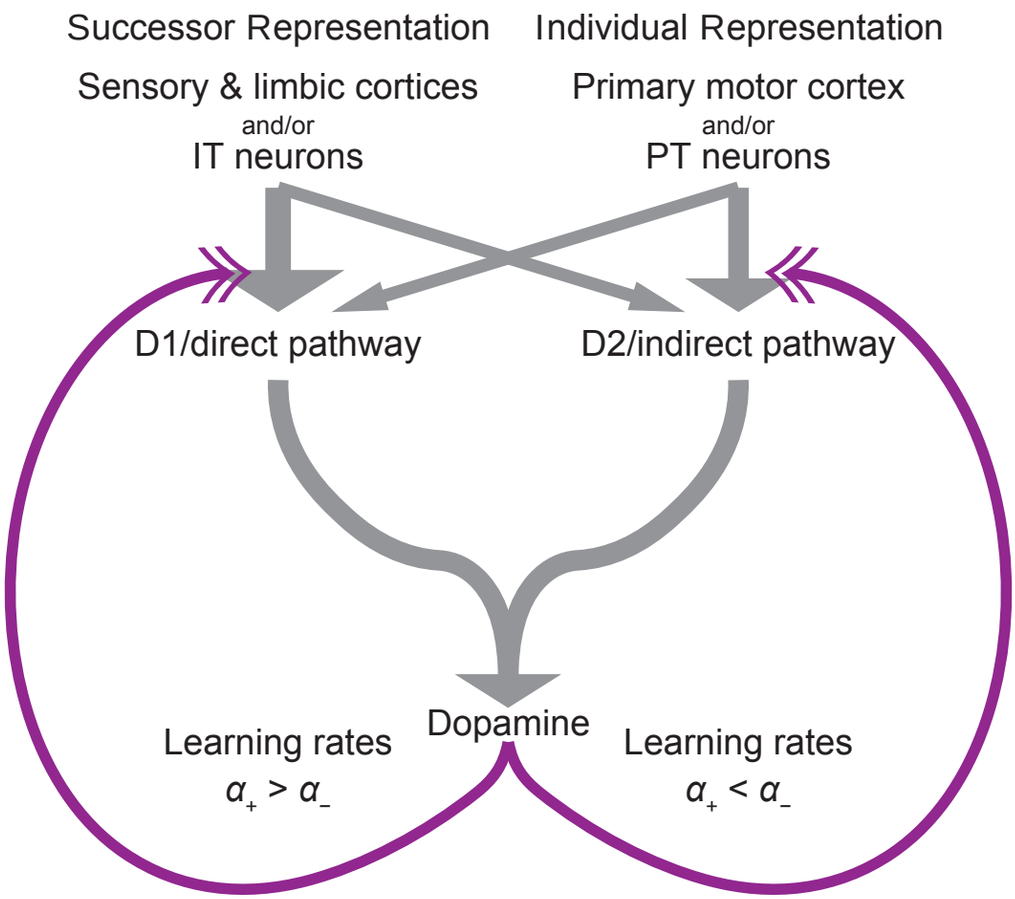

Successor Representation
Sensory & limbic cortices
and/or
IT neurons
Individual Representation
Primary motor cortex
and/or
PT neurons
D1/direct pathway
D2/indirect pathway
Dopamine
Learning rates
$\alpha_+ > \alpha_-$
Learning rates
$\alpha_+ < \alpha_-$

Figure 11

Table 1

SR & IR

γ	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
0.6	5	0.05	0.8	0.2	0.5	0.5	123.28
		0.1	0.8	0.2	0.35	0.35	123.16
		0.1	0.8	0.2	0.5	0.5	123
		0.1	0.65	0.2	0.35	0.5	122.34
		0.1	0.8	0.2	0.2	0.65	121.89
		0.1	0.8	0.2	0.5	0.8	121.59
		0.1	0.65	0.2	0.35	0.35	121.57
		0.1	0.8	0.2	0.35	0.5	121.45
		0.1	0.8	0.2	0.2	0.8	121.42
		0.05	0.8	0.2	0.5	0.65	121.3
	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
	10	0.05	0.65	0.2	0.2	0.65	127.94
		0.05	0.8	0.2	0.5	0.8	127.66
		0.05	0.65	0.2	0.2	0.8	127.36
		0.05	0.5	0.2	0.2	0.8	126.43
		0.05	0.8	0.2	0.65	0.8	126.12
		0.05	0.65	0.2	0.35	0.8	126.02
		0.05	0.65	0.2	0.35	0.65	125.89
		0.05	0.8	0.2	0.2	0.8	125.71
		0.05	0.5	0.2	0.35	0.8	125.63
		0.05	0.65	0.35	0.2	0.8	125.23
	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
	15	0.05	0.5	0.2	0.35	0.8	127.02
		0.05	0.5	0.2	0.2	0.8	125.85
		0.05	0.65	0.2	0.2	0.8	125.78
		0.05	0.65	0.2	0.65	0.8	125.75
		0.05	0.65	0.35	0.5	0.8	125.31
		0.05	0.65	0.2	0.2	0.65	124.71
		0.05	0.8	0.35	0.5	0.8	124.66
		0.05	0.65	0.2	0.35	0.8	124.22
		0.05	0.8	0.35	0.2	0.8	123.99
		0.05	0.8	0.8	0.2	0.8	123.91

SR in both

γ	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
0.6	5	0.15	0.5	0.2	0.5	0.2	122.12
		0.2	0.65	0.35	0.5	0.2	121.65
		0.15	0.8	0.2	0.65	0.2	121.05
		0.25	0.35	0.35	0.8	0.2	120.93
		0.2	0.65	0.2	0.65	0.2	120.89
		0.2	0.5	0.2	0.35	0.2	120.79
		0.15	0.65	0.35	0.65	0.2	120.67
		0.15	0.8	0.2	0.5	0.2	120.54
		0.15	0.8	0.2	0.2	0.2	120.47
		0.1	0.8	0.35	0.2	0.2	120.47
	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
	10	0.2	0.5	0.35	0.2	0.35	123.04
		0.2	0.8	0.5	0.5	0.5	122.86
		0.2	0.5	0.35	0.2	0.2	122.84
		0.15	0.2	0.65	0.5	0.2	122.79
		0.25	0.2	0.5	0.8	0.2	122.56
		0.25	0.8	0.5	0.2	0.35	122.54
		0.15	0.35	0.8	0.8	0.35	122.29
		0.1	0.5	0.35	0.35	0.2	122.28
		0.2	0.65	0.65	0.8	0.5	122.24
		0.2	0.35	0.5	0.35	0.35	122.05
	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
	15	0.2	0.5	0.8	0.35	0.5	123.66
		0.15	0.65	0.65	0.35	0.35	123.47
		0.2	0.35	0.5	0.2	0.2	122.69
		0.15	0.2	0.8	0.5	0.35	122.68
		0.2	0.35	0.8	0.8	0.5	122.62
		0.1	0.5	0.65	0.2	0.2	122.48
		0.15	0.5	0.8	0.2	0.2	122.43
		0.15	0.35	0.8	0.65	0.35	122.35
		0.15	0.2	0.65	0.35	0.35	122.22
		0.1	0.8	0.65	0.35	0.2	122.21

IR in both

γ	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
0.6	5	0.65	0.2	0.5	0.2	110.14
		0.5	0.2	0.5	0.2	109.64
		0.65	0.2	0.35	0.2	109.07
		0.65	0.2	0.2	0.2	108.3
		0.65	0.35	0.5	0.2	108.29
		0.8	0.2	0.2	0.2	108.1
		0.8	0.2	0.35	0.2	108.08
		0.65	0.2	0.65	0.2	107.97
		0.5	0.5	0.8	0.2	107.96
		0.5	0.35	0.8	0.2	107.67
	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
	10	0.2	0.35	0.8	0.2	107.02
		0.2	0.35	0.35	0.2	106.23
		0.35	0.35	0.35	0.2	106.06
		0.2	0.35	0.65	0.2	105.99
		0.5	0.2	0.35	0.2	105.91
		0.5	0.35	0.2	0.2	105.88
		0.65	0.35	0.2	0.2	105.76
		0.2	0.35	0.5	0.2	105.4
		0.2	0.5	0.5	0.2	105.31
		0.35	0.35	0.2	0.35	105.31
	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
	15	0.2	0.2	0.2	0.2	105.81
		0.35	0.2	0.2	0.2	105.34
		0.5	0.35	0.2	0.2	105.14
		0.5	0.35	0.2	0.35	104.42
		0.2	0.35	0.35	0.2	103.65
		0.5	0.5	0.2	0.35	103.6
		0.2	0.5	0.5	0.2	103.59
		0.35	0.5	0.35	0.35	103.43
		0.2	0.5	0.35	0.35	103.24
		0.2	0.5	0.2	0.2	103.23

γ	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
0.7	5	0.1	0.8	0.2	0.35	0.5	125.6
		0.1	0.8	0.2	0.2	0.5	125.25
		0.1	0.65	0.2	0.2	0.8	124.28
		0.05	0.8	0.2	0.35	0.8	124.07
		0.1	0.65	0.2	0.65	0.65	123.93
		0.15	0.65	0.2	0.35	0.2	123.6
		0.1	0.65	0.2	0.65	0.8	123.6
		0.1	0.65	0.2	0.35	0.65	123.57
		0.1	0.5	0.2	0.35	0.65	123.51
		0.1	0.8	0.35	0.35	0.5	123.3
	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
	10	0.05	0.65	0.2	0.2	0.8	131.76
		0.05	0.8	0.2	0.2	0.8	128.7
		0.05	0.5	0.2	0.35	0.8	128.34
		0.05	0.35	0.2	0.2	0.8	128.15
		0.05	0.35	0.2	0.35	0.8	127.61
		0.05	0.8	0.2	0.35	0.8	127.36
		0.1	0.5	0.35	0.5	0.8	127.36
		0.05	0.5	0.2	0.2	0.5	127.35
		0.05	0.5	0.2	0.2	0.8	127.15
		0.1	0.5	0.2	0.65	0.8	126.98
	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
	15	0.05	0.35	0.2	0.35	0.8	126.89
		0.05	0.5	0.2	0.35	0.8	125.68
		0.05	0.5	0.2	0.2	0.8	125.36
		0.05	0.35	0.2	0.2	0.8	125.33
		0.1	0.2	0.2	0.35	0.8	125.01
		0.05	0.2	0.2	0.2	0.8	124.78
		0.05	0.5	0.2	0.5	0.8	124.66
		0.05	0.35	0.2	0.2	0.65	124.31
		0.05	0.8	0.35	0.5	0.8	124.21
		0.1	0.2	0.2	0.35	0.5	124.12

γ	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
0.7	5	0.1	0.8	0.2	0.5	0.2	121.1
		0.1	0.65	0.35	0.5	0.2	120.89
		0.2	0.5	0.2	0.35	0.2	120.83
		0.2	0.2	0.35	0.65	0.2	120.77
		0.1	0.65	0.2	0.2	0.2	120.61
		0.1	0.5	0.2	0.2	0.2	120.46
		0.1	0.35	0.2	0.35	0.2	120.23
		0.15	0.5	0.2	0.35	0.2	120.19
		0.15	0.2	0.35	0.65	0.2	120.19
		0.2	0.8	0.2	0.2	0.2	120.15
	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
	10	0.2	0.2	0.2	0.2	0.2	123.66
		0.15	0.2	0.65	0.65	0.2	123.59
		0.15	0.35	0.5	0.65	0.35	123.37
		0.15	0.35	0.8	0.35	0.2	122.75
		0.1	0.2	0.65	0.5	0.2	122.74
		0.1	0.5	0.65	0.35	0.35	122.57
		0.1	0.2	0.2	0.2	0.2	122.45
		0.1	0.8	0.5	0.2	0.5	122.38
		0.15	0.8	0.5	0.2	0.5	122.27
		0.1	0.5	0.65	0.35	0.5	122.19
	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
	15	0.15	0.35	0.5	0.2	0.35	122.25
		0.1	0.5	0.5	0.2	0.5	121.4
		0.1	0.2	0.8	0.2	0.2	121.22
		0.15	0.2	0.65	0.35	0.2	121.2
		0.1	0.2	0.65	0.35	0.35	120.98
		0.15	0.2	0.35	0.2	0.35	120.95
		0.15	0.2	0.5	0.65	0.35	120.95
		0.15	0.35	0.65	0.5	0.35	120.77
		0.1	0.35	0.65	0.2	0.35	120.72
		0.1	0.35	0.8	0.35	0.5	120.64

γ	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
0.7	5	0.8	0.2	0.2	0.2	109.71
		0.5	0.2	0.2	0.2	109.19
		0.8	0.2	0.35	0.2	109.18
		0.8	0.35	0.2	0.2	108.99
		0.2	0.35	0.65	0.2	108.84
		0.8	0.5	0.5	0.2	108.6
		0.65	0.35	0.65	0.2	108.53
		0.5	0.2	0.35	0.2	108.45
		0.65	0.2	0.35	0.2	108.45
		0.65	0.2	0.65	0.2	108.4
	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
	10	0.5	0.2	0.2	0.2	106.11
		0.5	0.5	0.2	0.2	106.05
		0.2	0.35	0.65	0.2	105.81
		0.35	0.35	0.35	0.2	105.79
		0.35	0.35	0.2	0.2	105.76
		0.35	0.2	0.2	0.2	105.52
		0.2	0.2	0.2	0.2	105.3
		0.5	0.35	0.35	0.2	105.28
		0.2	0.5	0.5	0.35	105.26
		0.2	0.5	0.5	0.2	105.22
	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
	15	0.35	0.5	0.2	0.2	106
		0.2	0.35	0.2	0.35	105.58
		0.2	0.35	0.2	0.2	104.81
		0.2	0.2	0.2	0.2	104.18
		0.2	0.5	0.2	0.35	104.08
		0.2	0.65	0.2	0.5	103.64
		0.35	0.5	0.2	0.35	103.57
		0.35	0.35	0.2	0.35	103.56
		0.5	0.5	0.2	0.5	103.2
		0.35	0.35	0.35	0.35	103.19

γ	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
0.8	5	0.1	0.5	0.2	0.5	0.8	123.41
		0.1	0.35	0.2	0.5	0.65	122.58
		0.1	0.8	0.5	0.35	0.8	122.51
		0.1	0.5	0.35	0.2	0.8	122.27
		0.05	0.65	0.2	0.5	0.8	122.16
		0.1	0.65	0.2	0.5	0.65	121.81
		0.1	0.35	0.2	0.2	0.5	121.75
		0.1	0.35	0.2	0.35	0.65	121.4
		0.1	0.65	0.35	0.2	0.8	121.36
		0.15	0.5	0.35	0.5	0.8	121.29
	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
	10	0.1	0.35	0.2	0.35	0.8	126.11
		0.1	0.35	0.35	0.35	0.8	125.9
		0.1	0.2	0.2	0.2	0.8	125.59
		0.05	0.35	0.2	0.35	0.8	125.35
		0.05	0.5	0.35	0.2	0.8	125.25
		0.1	0.35	0.35	0.2	0.35	125.18
		0.1	0.35	0.35	0.2	0.65	125.11
		0.1	0.2	0.2	0.2	0.65	124.84
		0.15	0.35	0.35	0.35	0.8	124.83
		0.05	0.65	0.5	0.2	0.8	124.72
	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
	15	0.05	0.2	0.2	0.2	0.8	125.62
		0.05	0.35	0.35	0.2	0.8	124.09
		0.15	0.2	0.5	0.2	0.65	124.08
		0.05	0.2	0.2	0.2	0.65	123.97
		0.05	0.35	0.2	0.2	0.8	123.01
		0.15	0.2	0.2	0.35	0.5	122.85
		0.05	0.2	0.2	0.35	0.8	122.6
		0.2	0.2	0.35	0.2	0.35	122.34
		0.05	0.35	0.35	0.35	0.8	122.14
		0.1	0.2	0.2	0.5	0.8	121.93

γ	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
0.8	5	0.15	0.5	0.35	0.2	0.2	120.31
		0.2	0.2	0.35	0.35	0.2	120.24
		0.1	0.2	0.35	0.5	0.2	119.9
		0.1	0.35	0.35	0.5	0.2	119.9
		0.1	0.2	0.35	0.35	0.2	119.29
		0.1	0.35	0.35	0.35	0.2	118.75
		0.1	0.35	0.5	0.5	0.2	118.51
		0.1	0.35	0.35	0.35	0.35	118.28
		0.1	0.5	0.65	0.2	0.2	118.26
		0.1	0.35	0.65	0.5	0.2	118.01
	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
	10	0.1	0.2	0.35	0.2	0.35	121.9
		0.15	0.35	0.35	0.2	0.35	121.61
		0.15	0.2	0.35	0.2	0.2	121.39
		0.15	0.2	0.35	0.2	0.35	121.39
		0.1	0.5	0.35	0.2	0.35	120.83
		0.15	0.2	0.5	0.5	0.35	120.48
		0.15	0.2	0.65	0.5	0.2	120.21
		0.1	0.2	0.2	0.2	0.2	120.01
		0.15	0.35	0.5	0.2	0.2	120.01
		0.15	0.2	0.65	0.2	0.35	119.92
	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
	15	0.05	0.2	0.65	0.35	0.2	118.64
		0.05	0.2	0.65	0.2	0.35	118.18
		0.1	0.35	0.65	0.2	0.35	118.16
		0.05	0.35	0.5	0.2	0.5	117.88
		0.05	0.2	0.5	0.2	0.2	117.59
		0.05	0.2	0.5	0.2	0.35	117.52
		0.1	0.35	0.5	0.2	0.35	117.37
		0.1	0.35	0.65	0.2	0.2	117.15
		0.05	0.2	0.5	0.2	0.5	117.12
		0.1	0.35	0.8	0.2	0.2	117.08

γ	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
0.8	5	0.35	0.35	0.8	0.2	109.55
		0.35	0.35	0.5	0.2	109.24
		0.2	0.65	0.8	0.2	109.24
		0.5	0.2	0.5	0.2	109.22
		0.8	0.2	0.35	0.2	109.19
		0.65	0.35	0.35	0.2	108.86
		0.65	0.2	0.5	0.2	108.83
		0.65	0.2	0.2	0.2	108.77
		0.65	0.5	0.35	0.2	108.6
		0.35	0.5	0.8	0.35	108.51
	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
	10	0.35	0.5	0.2	0.2	107.23
		0.2	0.35	0.35	0.2	106.89
		0.2	0.2	0.2	0.2	106.45
		0.2	0.65	0.5	0.35	106.1
		0.5	0.65	0.2	0.2	106.04
		0.2	0.5	0.35	0.35	105.99
		0.2	0.35	0.2	0.35	105.97
		0.2	0.35	0.2	0.2	105.92
		0.2	0.8	0.35	0.35	105.85
		0.35	0.35	0.2	0.2	105.78
	β	IR1 α_+	α_-	IR2 α_+	α_-	Reward
	15	0.2	0.2	0.2	0.2	106.84
		0.2	0.35	0.2	0.2	105.63
		0.2	0.65	0.35	0.2	105.45
		0.2	0.35	0.2	0.35	105.42
		0.2	0.65	0.2	0.2	104.5
		0.2	0.5	0.35	0.35	104.2
		0.35	0.5	0.2	0.2	103.95
		0.35	0.35	0.2	0.35	103.83
		0.2	0.5	0.35	0.2	103.76
		0.2	0.65	0.2	0.35	103.76

Table 2

SR in both

γ	β	$\alpha_{feature}$	SR1 α_+	α_-	SR2 α_+	α_-	Reward
0.5	15	0.2	0.35	0.35	0.35	0.35	123.65
		0.1	0.8	0.8	0.35	0.35	123.04
		0.15	0.35	0.65	0.5	0.5	122.88
		0.25	0.5	0.65	0.8	0.5	122.88
		0.15	0.5	0.5	0.35	0.2	122.79
		0.1	0.65	0.8	0.35	0.2	122.65
		0.2	0.65	0.65	0.8	0.35	122.48
		0.15	0.35	0.5	0.35	0.35	122.28
		0.25	0.5	0.8	0.8	0.5	122.25
		0.2	0.65	0.8	0.5	0.2	122.17

γ	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
0.5	20	0.2	0.2	0.8	0.8	0.2	123.58
		0.25	0.5	0.65	0.65	0.2	123.53
		0.2	0.65	0.5	0.2	0.5	123.09
		0.2	0.8	0.65	0.2	0.2	122.99
		0.15	0.5	0.8	0.65	0.5	122.33
		0.2	0.65	0.65	0.2	0.2	122.28
		0.2	0.2	0.65	0.65	0.35	122.26
		0.2	0.5	0.8	0.5	0.5	122.2
		0.25	0.2	0.65	0.35	0.35	122.03
		0.25	0.8	0.8	0.2	0.5	122.02

γ	β	$\alpha_{feature}$	SR α_+	α_-	IR α_+	α_-	Reward
0.6	20	0.15	0.65	0.5	0.2	0.5	124
		0.1	0.2	0.5	0.5	0.35	121.91
		0.15	0.35	0.8	0.2	0.2	121.71
		0.2	0.2	0.65	0.8	0.5	121.66
		0.15	0.2	0.5	0.2	0.2	121.59
		0.05	0.2	0.65	0.35	0.5	121.46
		0.1	0.2	0.35	0.2	0.35	121.4
		0.1	0.35	0.8	0.2	0.35	121.33
		0.1	0.2	0.65	0.8	0.2	121.31
		0.2	0.2	0.35	0.2	0.35	121.31