Title

Opponent Learning with Different Representations in the Cortico-Basal Ganglia Pathways Can Develop Obsession-Compulsion Cycle

Author names and affiliations

Reo Sato[1+], Kanji Shimomura[1+], & Kenji Morita[1,2*]

[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

[+] These authors contributed equally to this work.

**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 that combined SR- and IR-based systems could develop OCD-like behavior, and elaborated the idea in discussion with RS and KS. RS, KS, and KM conducted/validated the simulations and analyses. KM drafted the manuscript, and revised it with inputs from RS and KS. According to the CRediT Taxonomy, Conceptualization: KM; Investigation: RS, KS, KM; Funding Acquisition: KM; Writing – Original Draft Preparation: KM; Writing – Review & Editing: RS, KS, KM.

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 (MEXT) (http://www.mext.go.jp/en/) and the Japan Society for the Promotion of Science (JSPS) (https://www.jsps.go.jp/english/) and the Naito Foundation (https://www.naito-f.or.jp/en/).

Conflict of interest

None

Abstract

Obsessive-compulsive disorder (OCD) has been suggested to be associated with impairment of model-based behavioral control. Meanwhile, recent work suggested shorter memory trace for negative than positive prediction errors (PEs) in OCD. Relation between these two suggestions remains unclear. Based on the properties of cortico-basal ganglia pathways, we modeled human as agent having a combination of successor representation (SR)-based system that enables model-based-like control and individual representation (IR)-based system that only hosts model-free control. We show that if the SR- and IR-based systems learn mainly from positive and negative PEs, respectively, agent can develop obsession-compulsion cycle, similarly to the agent with memory trace imbalance in the previous study. Fitting of behavior of such an opponent SR+IR agent in the two-stage decision task resulted in smaller weights for model-based control than the SR-only agent. These results reconcile the previous suggestions about OCD, raising a possibility that opponent learning in model-based/free controls underlies obsession-compulsion. Moreover, together with a recent suggestion that the opponent SR+IR combination actually performs well in certain dynamic environments and its cortico-striatal implementation appears consistent with various findings, this possibility explains why human is prone to OCD and even healthy people tended to show shorter memory trace for negative PEs.

Introduction

Obsessive-compulsive disorder (OCD) has been suggested to be associated with impairment of model-based/goal-directed behavioral control, as compared to model-free (habitual) control (Robbins et al., 2019). In particular, analysis of behavioral choices of general populations in the two-stage decision task (Gillan et al., 2016) revealed that a symptom dimension "compulsive behavior and intrusive thought", which could exist in OCD and also in other disorders such as addiction, was strongly associated with deficits in model-based control. Meanwhile, recent work (Sakai et al., 2022) suggested that memory trace imbalance underlies OCD. Specifically, modeling showed that obsession-compulsion cycle can be enhanced in agent having much shorter memory (eligibility) trace for negative than positive prediction errors (PEs), and fitting of behavioral choices suggested that OCD patients indeed had such an imbalance as compared to healthy controls (HCs), although even HCs tended to show shorter memory trace for negative than positive PEs. Mechanisms of such memory trace imbalance, as well as its relation to the suggestion of impaired model-based control, remain unclear.

Whereas the abovementioned modeling in the recent work (Sakai et al., 2022) examined agent that only had model-free control, multiple studies have suggested that humans and other animals can use both model-based and model-free controls (Dolan and Dayan, 2013). Model-free control has been suggested to be implemented in the way that dopamine represents reward PEs (RPEs) and dopamine-dependent plasticity of cortico-striatal synapses represents RPE-based update of state/action values (Montague et al., 1996; Schultz et al., 1997). Neural implementation of model-based control has remained more elusive, but recent studies suggest that if states or actions are represented in a manner that contains information about action-state transitions, partially model-based-like control can be acquired through RPE-based update and thereby through dopamine-dependent plasticity, similarly to model-free control (Russek et al., 2017). As such a representation, successor representation (SR), in which state/action is represented by future occupancies of successor states/actions (Dayan, 1993), has been suggested to be used in humans (Momennejad et al., 2017) and implemented in the brain (Garvert et al., 2017; Stachenfeld et al., 2017; Russek et al., 2021).

Different cortical regions or neural populations may adopt different ways of representations, in particular, SR and more classical individual (punctate) representation (IR), and they may unevenly project to the direct and indirect pathways of basal ganglia (Lei et al., 2004; Wall et al., 2013; Lu et al., 2021), which have been suggested to be crucial for learning from positive and negative feedbacks, respectively (Frank et al., 2004; Hikida et al., 2010; Kravitz et al., 2012; Tai et al., 2012; Iino et al., 2020; Lee et al., 2021). Different circuits for positive and negative learning other than these pathways have also been suggested (Skvortsova et al., 2014; Groman et al., 2019). Given these, it is conceivable that SR-based system and IR-based system differentially learn from positive and negative RPEs in the brain. Recent work (Morita et al., 2022) explored this possibility, and found that combination of SR-

and IR-based systems learning mainly from positives and negative RPEs, respectively (referred to as appetitive SR- and aversive IR-based systems), performed well in certain dynamic reward environments. This work further suggested that implementation of such a combination in the cortico-basal ganglia pathways seems in line with various anatomical and physiological findings.

Crucially, SR has a similarity to the eligibility trace (c.f., its "forward view" (Sutton and Barto, 2018)) while model-free learning without eligibility trace is naturally described by using IR. Given this correspondence, we expected that the combination of appetitive SR- and aversive IR-based systems could potentially show similar behavior to the one shown by the agent having long / short eligibility traces for positive/ negative RPEs in the abovementioned recent work (Sakai et al., 2022), in particular, enhancement of obsession-compulsion cycle. Moreover, we also expected that the combined appetitive SR- and aversive IR-based systems could potentially behave in a less-model-based manner than a fully model-based or SR-based system in the two-stage decision task (Daw et al., 2011), whereas eligibility-trace imbalance *per se* would not bias the estimation of the degree of model-based control. Here we addressed these expectations.

Results

We adopted the previously developed environmental model that describes possible enhancement of obsession-compulsion cycle (Sakai et al., 2022) (Figure 1). There are two states named the relief state and the obsession state. At the relief state, agent can take one of two actions named the "abnormal reaction" and the "other" action. Here it is presumed that occasionally some anxiety e.g., about whether the door was locked arises spontaneously (intrusive thought), and agent can either get deeply into the anxiety ("abnormal reaction"), entering the obsession state, or just ignore or forget about it and do other things ("other" action), remaining in the relief state. Once agent enters the obsession state, agent again can take one of two actions. If agent takes the "compulsion" action by paying a small cost, e.g., confirms door lock, agent can transition back to the relief state with a high probability. Alternatively, agent can take "other" action without any cost, but it makes the agent transition back to the relief state only with a small probability. Every stay at the obsession state imposes punishment (negative reward).

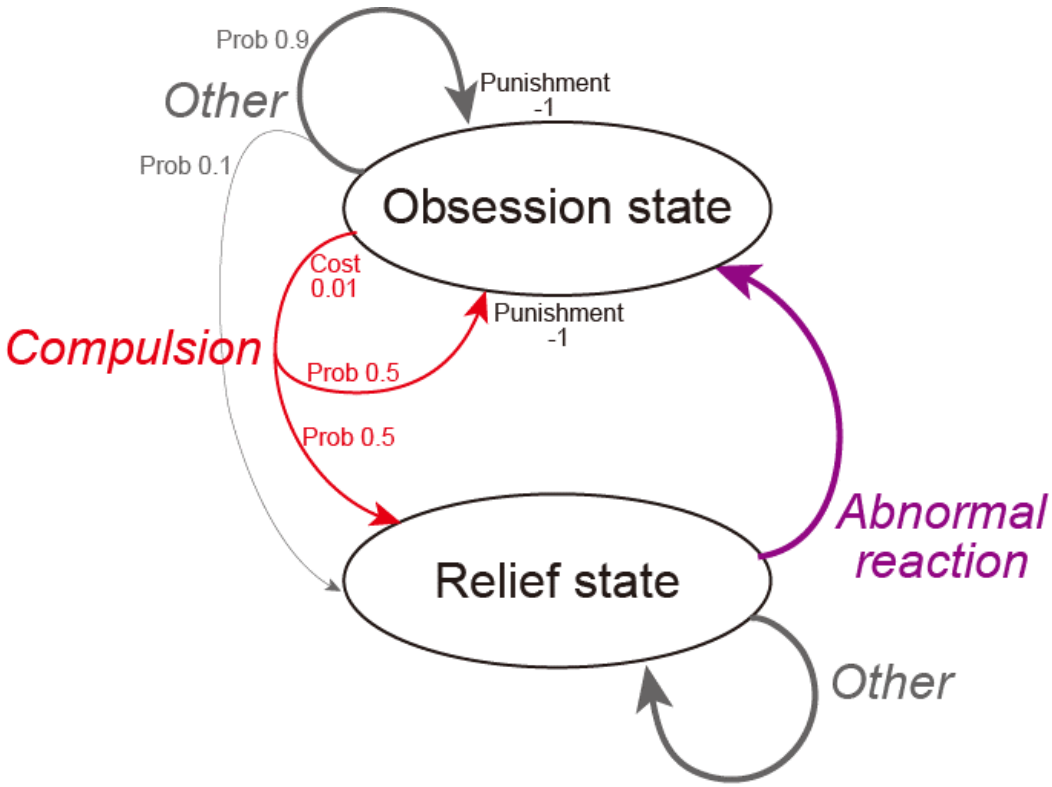

Figure 1

Figure 1
Environmental model describing possible development of obsession-compulsion cycle, adopted from the previous study (Sakai et al., 2022). There are two states: the relief state and the obsession state. At the relief state, agent can take the "abnormal reaction" (to an intrusive thought, i.e., spontaneously arising anxiety e.g., about door lock), which induces a transition to the obsession state, or the "other" action (e.g., just ignore or forget about it), with which the agent stays at the relief state. At the obsession state, agent can take the "compulsion" action (e.g., confirms door lock), which requires a small cost (0.01) but induces a transition back to the relief state with a high probability (50%), or the "other" action, which requires no cost but induces a transition back to the relief state only with a small probability (10%). Every stay at the obsession state imposes punishment (negative reward −1).

Because of this punishment, the optimal value of the obsession state should be negative and lower than the optimal value of the relief state. Thus, whereas the "compulsion" action can have a positive value because it may cause transition from the lower valued obsession state to the higher valued relief state, the optimal value of the "abnormal reaction", which induces transition to the obsession state, should be negative and lower than the optimal value of the "other" action at the relief state. Therefore, normative reinforcement learning (RL) agent should learn a policy that minimizes selection of the "abnormal reaction", thereby never developing obsession-compulsion cycle. However, agent that deviates from normative RL, such as the one having different lengths of eligibility traces for positive versus negative RPEs examined in the previous work (Sakai et al., 2022), may behave differently.

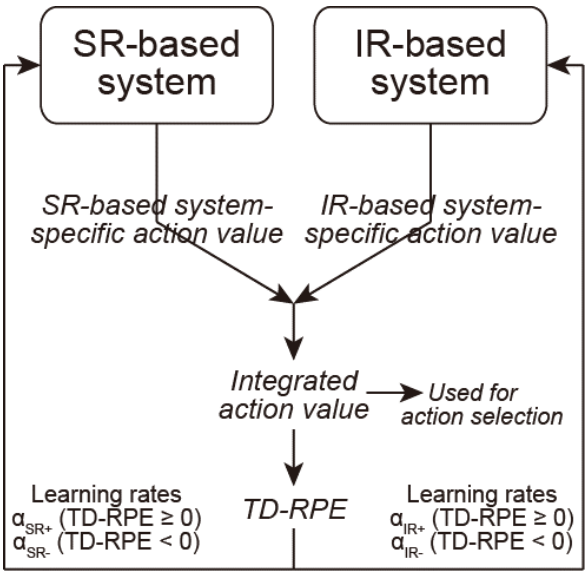

Figure 2

Figure 2
Agent having a coupled SR-based system and IR-based system, adopted from the previous study (Morita et al., 2022). Each system develops the system-specific value of each action, and the average of the two system-specific values, named the integrated action value, is used for soft-max action selection and calculation of SARSA-type TD-RPEs. The TD-RPEs are used for updating the system-specific values, with the learning rates for the SR- and IR-based systems in the cases of positive (non-negative) and negative TD-RPEs, denoted as α_{SR+}, α_{SR-}, α_{IR+}, and α_{IR-}, can take different values.

We examined how agent having coupled SR- and IR-based systems (Figure 2) behaved in this environmental model. First, we confirmed whether agent that effectively had only the IR-based system with the same learning rates from positive and negative RPEs (denoted as α_{IR+} and α_{IR-}), i.e., TD(0) agent, behaved in the normative manner as described above, by setting the learning rates of the SR-based system from positive and negative RPEs (α_{SR+} and α_{SR-}) to 0 while α_{IR+} and α_{IR-} were set to 0.1. Figure 3A-top shows an example of the time evolution of the difference in the action values of "abnormal reaction" and "other" at the relief state ($Q_{abnormal} - Q_{other@relief}$). Figure 3A-middle shows the moving average of the proportion that the agent was at the obsession state in the same example simulation, and Figure 3A-bottom shows the average of such a moving-average proportion across 100 simulations. As shown in these figures, it was confirmed that this agent minimized selection of the "abnormal reaction" at the relief state and never developed obsession-compulsion cycle. We next examined the case where agent effectively had only the SR-based system with unbiased learning from positive and negative PREs, by setting α_{IR+} and α_{IR-} to 0 while α_{SR+} and α_{SR-} were set to 0.1. As shown in Figure 3B, this agent also minimized selection of the "abnormal reaction". This is actually reasonable given that SR-based temporal-difference (TD) leaning can generally approximate optimal value functions.

Next, we examined the behavior of agent in which both IR-based system and SR-based system were effective and had the same learning rates for both systems and also from positive and negative RPEs (i.e., $\alpha_{IR+} = \alpha_{IR-} = \alpha_{SR+} = \alpha_{SR-} = 0.05$). As shown in Figure 3C, this agent also minimized selection of the "abnormal reaction". This may also be reasonable, as this agent combined rigorous normative TD(0) leaner and approximation of normative TD leaner. We then examined the behavior of agent having appetitive SR- and aversive IR-based systems by setting $\alpha_{SR+} = \alpha_{IR-} = 0.09$ and $\alpha_{SR-} = \alpha_{IR+} = 0.01$. Figure 3D show the results. As shown in Figure 3D-top, different from the previous cases, the $Q_{abnormal} - Q_{other@relief}$ difference did not smoothly decrease but remained to fluctuate and sometimes even became positive, indicating that this agent continued to take the "abnormal reaction" and transition to the obsession state rather frequently, as actually appeared in Figure 3D-middle,bottom. In other words, obsession-compulsion cycle was often developed and enhanced in this agent, as we expected. We further examined cases with different learning rates while $\alpha_{SR+} + \alpha_{IR+}$ and $\alpha_{SR-} + \alpha_{IR-}$ (i.e., total learning rates from positive and negative RPEs, respectively) were kept constant at 0.1,

finding that agent with large α_{SR+} & small α_{IR+} and small α_{SR-} & large α_{IR-} (i.e., appetitive SR- and aversive IR-based systems) could develop obsession-compulsion cycle (Figure 3E). In the long run, such a cycle ultimately terminated (Figure 3F). However, if the rate of the update of SR matrix, which was fixed at 0.01 in the simulations shown so far, decreased over time and converged to 0, i.e., SR became rigid (c.f., (Shimomura et al., 2021)), obsession-compulsion cycle appeared to persist (Figure 3G).

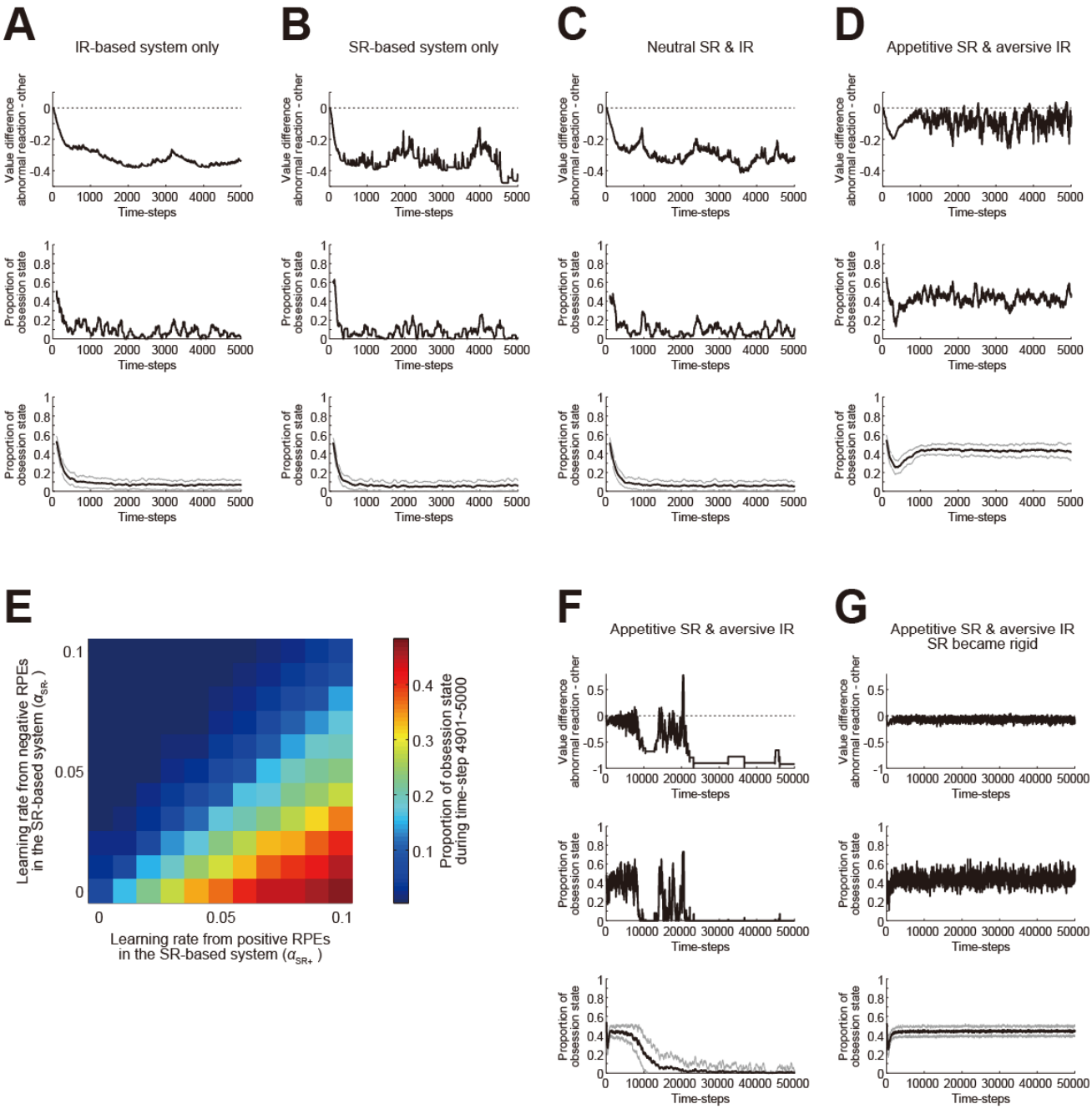

Figure 3

Figure 3
Behavior of different types of agent. **(A)** Behavior of agent that effectively had IR-based system only ((α_{SR+}, α_{SR-}, α_{IR+}, α_{IR-}) = (0, 0, 0.1, 0.1)). *Top*: An example of the time evolution of the difference in the action values of the "abnormal reaction" and the "other" at the relief state ($Q_{abnormal} - Q_{other@relief}$). *Middle*: The moving average (over the past 100 time steps) of the proportion that the agent was at the obsession state in the same example simulation. *Bottom*: The average of the moving-average proportion of the obsession state across 100 simulations (black line), presented with ±SD (gray thin lines). **(B)** Behavior of agent that effectively had SR-based system only ((α_{SR+}, α_{SR-}, α_{IR+}, α_{IR-}) = (0.1, 0.1, 0, 0)). **(C)** Behavior of agent having IR- and SR-based systems, both of which learned equally from positive and negative RPEs ((α_{SR+}, α_{SR-}, α_{IR+}, α_{IR-}) = (0.05, 0.05, 0.05, 0.05)). **(D)** Behavior of agent having appetitive SR- and aversive IR-based systems ((α_{SR+}, α_{SR-}, α_{IR+}, α_{IR-}) = (0.09, 0.01, 0.01, 0.09)). **(E)** The average of the moving-average proportion of the obsession state at time-step 5000 (i.e., moving-average during time-step 4901~5000) across 100 simulations in various cases with different learning rates. The horizontal and vertical axes indicate the learning rates of the SR-based system from positive and negative RPEs (i.e., α_{SR+} and α_{SR-}), respectively, while $\alpha_{SR+} + \alpha_{IR+}$ and $\alpha_{SR-} + \alpha_{IR-}$ (i.e., total learning rates from positive and negative RPEs, respectively) were kept constant at 0.1. **(F,G)** Behavior of the agent with appetitive SR- and aversive IR-based systems in the long run, i.e., for 50000 time steps, in the cases where the rate of the update of SR matrix was fixed at 0.01 (F) and it decreased over time according to 0.01/(1 + time-step/1000) (G).

We examined how the agent with appetitive SR- and aversive IR-based systems behaved in the two-stage decision task (Daw et al., 2011), in comparison to the agent with SR-based system only that would correspond to healthy controls who had long memory traces for both positive and negative RPEs in the previous work (Sakai et al., 2022). Specifically, we simulated choice sequences of both agents 100 times for each using different pseudorandom numbers, and fitted them by the RL model weighing model-based and model-free controls developed in the original study (Daw et al., 2011) to estimate the parameter w that represents the degree of model-based control. As shown in Figure 4A,B, the distribution of the estimated w for the agent with appetitive SR- and aversive IR-based systems (Figure 4B) was smaller than that for the agent with SR-based system only (Figure 4A) (Wilcoxon rank sum test, $p = 1.04\times10^{-7}$). In contrast, fitting of choices generated by the original RL model with balanced eligibility trace ($\lambda = 0.4$ for both positive and negative RPEs) and those generated by a modified model with imbalance trace ($\lambda = 0.4$ and 0 for positive and negative RPEs, respectively), both of which had the same degree of model-based control ($w = 0.5$), did not result in smaller estimated w in the latter, as shown in Figure 4C,D (Wilcoxon rank sum test, $p = 0.5134$). Therefore, our expectation was again verified.

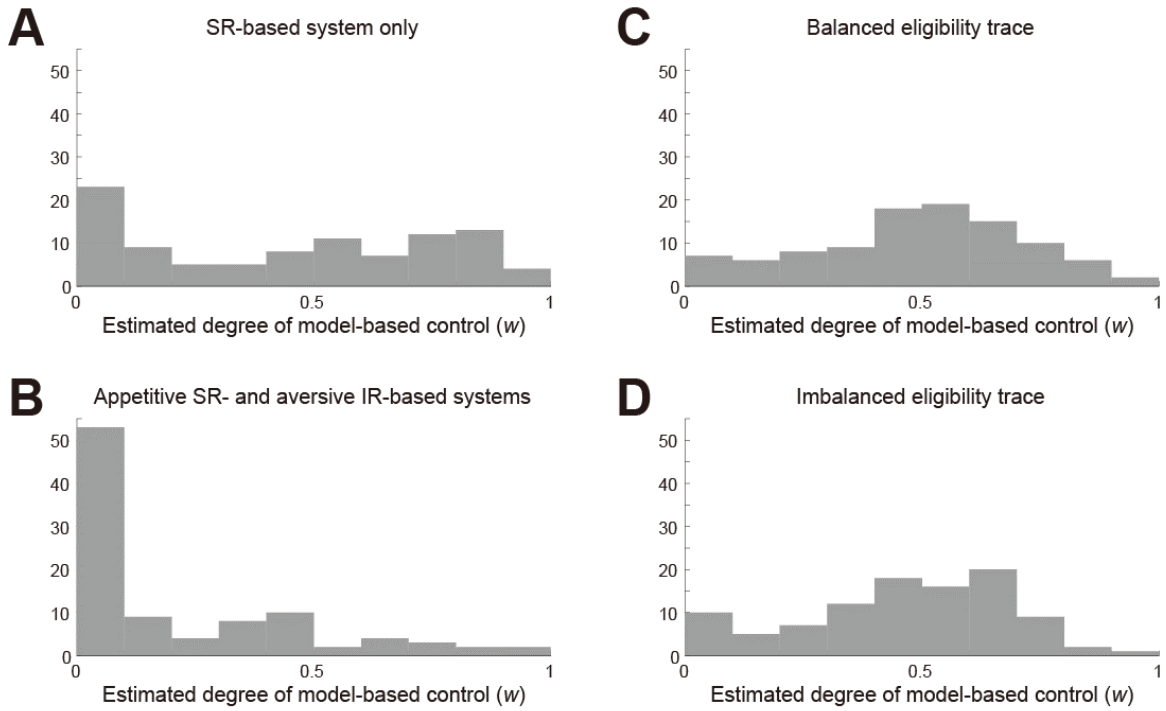

Figure 4

Figure 4

Estimation of the degree of model-based control from the agent's choice behavior in the two-stage decision task. **(A,B)** Comparison between the agent having SR-based system only (A) and the agent having appetitive SR- and aversive IR-based systems (B). The distributions of the estimated parameter representing the degree of model-based control (w) across simulations for each type of agent are shown. **(C,D)** Comparison between the RL model used in the original study with balanced eligibility trace (λ = 0.4) (C) and a modified model with imbalanced eligibility trace ((λ_1, λ_2) = (0.4, 0) for positive and negative TD-RPE, respectively) (D). The degree of model-based control (w) was set to 0.5 in the simulations of both models.

Discussion

These results reconcile the recent suggestion of memory trace imbalance in OCD (Sakai et al., 2022) with the long-standing suggestion that OCD is associated with impairment of model-based control (Gillan et al., 2016; Robbins et al., 2019), raising a possibility that opponent learning in model-based (SR-based) and model-free (IR-based) controls underlies obsession-compulsion. Moreover, given that such a combination of opponent SR+IR systems actually performs well in certain dynamic environments and its implementation in the cortico-basal ganglia pathways appears consistent with various physiological and anatomical findings (Morita et al., 2022), this possibility can further explain why human is prone to OCD and even healthy people tended to show shorter memory trace for negative PEs in the previous study (Sakai et al., 2022).

In addition, considering a recent finding that silencing of dorsal raphe serotonin neurons disrupts model-based control in outcome devaluation tasks (Ohmura et al., 2021), our model is in line with one of the suggested mechanisms of the effectiveness of selective serotonin reuptake inhibitors (SSRIs) for OCD that SSRIs re-engage the brain system for model-based control (Robbins et al., 2019). Moreover, impaired SR-based (model-based) learning from negative RPEs in OCD suggested by our model is potentially consistent with a previous result shown in (Worbe et al., 2016) that diminishing serotonin signaling by tryptophan depletion appeared to impair model-based choice after no-reward but not after reward (their Figure 2a, compare the left two panels) in the reward version of the task. In the punishment version of the task, however, tryptophan depletion rather promoted model-based behavior, and given that the environmental model that we used (developed by (Sakai et al., 2022)) only contained punishment and cost, this point appears inconsistent. Nonetheless, even if the environmental model was altered in a way that punishment (−1) for the stay at the obsession state was replaced with no reward for the stay at the obsession state and reward (1) for all the other cases, agent's behavior reported here was largely preserved provided that the IR-system-specific values and the weights for SR-system-specific values were initialized to $1+\gamma+\gamma^2+...=1/(1-\gamma)=2$ instead of 0. But still, the lack of outcome-valence dependence is an important limitation of our model (c.f., (Worbe et al., 2016) and also (Voon et al., 2015))).

Materials and Methods

Environmental model that describes possible enhancement of obsession-compulsion cycle

We adopted the environmental model developed by the previous study (Sakai et al., 2022) that describes possible enhancement of obsession-compulsion cycle (Figure 1). There are two states, i.e., the relief state and the obsession state, and two available actions at each state, i.e., the "abnormal reaction" and the "other" action at the relief state and the "compulsion" and the "other" action at the obsession state. State transitions are determined as follows: at the relief state, taking the abnormal reaction always causes transition to the obsession state whereas taking the other action always causes stay at the relief state, and at the obsession state, taking the compulsion causes transition to the relief state and stay at the obsession state with 50% probability each whereas taking the other action causes transition to the relief state with 10% probability and stay at the obsession state with 90% probability. Taking the compulsion at the obsession state requires cost (negative reward) −0.01, and every stay at the obsession state imposes punishment (negative reward) −1; no other reward is assumed. The parameter values regarding state transitions and rewards are the same as those used in the previous work (Sakai et al., 2022).

Agent having coupled SR- and IR-based systems

We adopted the model agent developed by the recent work (Morita et al., 2022) that have coupled SR- and IR-based systems (Figure 2), with a modification that SR of actions (c.f., (Russek et al., 2017)) rather than SR of states was used in the present study because there were only two states in the environmental model. Each system developed the system-specific value of each action, and their average, referred to as the (integrated) action value, was used for action selection and TD-RPE calculation.

Action selection was made in a soft-max manner. Specifically, when there were two available actions A_1 and A_2, whose action values were $Q(A_1)$ and $Q(A_2)$, A_1 was selected with probability $\exp(\beta Q(A_1)) / \{\exp(\beta Q(A_1)) + \exp(\beta Q(A_2))\}$ and A_2 was selected with the remaining probability, where β was the inverse temperature parameter representing the degree of choice exploitation over exploration and was set to 10.

SARSA-type TD-RPE was used for the results presented in the figures. Specifically, when agent took action A_{t-1}, payed cost c (0.01 for the compulsion and 0 for the other actions), received reward r (−1 if staying at the obsession state and 0 otherwise), transitioned to the next state, and took the next action A_t, SARSA-type TD-RPE $\delta = -c + r + \gamma Q(A_t) - Q(A_{t-1})$, where γ was the time discount factor and was set to 0.5, was calculated. We also examined the cases with Q-learning-type TD-RPE, and the main properties of agent's behavior described in the main text were not largely changed. Along

with TD-RPE, TD errors for SR features were calculated: $\mathbf{\Delta} = \mathbf{e}_{t-1} + \gamma\mathbf{M}_t, - \mathbf{M}_{t-1}$, where $\mathbf{e}_{t-1}$ was the vector that had four elements corresponding to the four actions and the element corresponding to A_{t-1} was 1 and all the other elements were 0, and $\mathbf{M}_i$ ($i = t-1$ or t) was the A_i-corresponding row of the SR matrix $\mathbf{M}$.

IR-based system-specific value of A_{t-1}, $Q_{\mathrm{IR}}(A_{t-1})$, was updated based on the TD-RPE: $Q_{\mathrm{IR}}(A_{t-1}) \leftarrow Q_{\mathrm{IR}}(A_{t-1}) + \alpha_{\mathrm{IR}+}\delta$ (when $\delta \geq 0$) or $Q_{\mathrm{IR}}(A_{t-1}) + \alpha_{\mathrm{IR}-}\delta$ (when $\delta < 0$), where $\alpha_{\mathrm{IR}+}$ and $\alpha_{\mathrm{IR}-}$ were the learning rates in the IR-based system for positive (non-negative) and negative TD-RPEs, respectively. SR-based system-specific value of A_{t-1}, $Q_{\mathrm{SR}}(A_{t-1})$, was given by the dot product of the SR feature vector for A_{t-1}, i.e., $\mathbf{M}_{t-1}$, and the weight vector for the approximate SR system-specific value function, which was common to all the actions and denoted by $\mathbf{w}$, i.e., $\mathbf{w}\cdot\mathbf{M}_{t-1}$. The weight $\mathbf{w}$ was updated based on the TD-RPE: $\mathbf{w} \leftarrow \mathbf{w} + \alpha_{\mathrm{SR}+}\delta\mathbf{M}_{t-1}$ (when $\delta \geq 0$) or $\mathbf{w} + \alpha_{\mathrm{SR}-}\delta\mathbf{M}_{t-1}$ (when $\delta < 0$), where $\alpha_{\mathrm{SR}+}$ and $\alpha_{\mathrm{SR}-}$ were the learning rates in the SR-based system for positive (non-negative) and negative TD-RPEs, respectively. The SR feature vector for A_{t-1}, $\mathbf{M}_{t-1}$, was updated based on the TD errors for SR features: $\mathbf{M}_{t-1} \leftarrow \mathbf{M}_{t-1} + \alpha_{\mathrm{feature}}\mathbf{\Delta}$, where $\alpha_{\mathrm{feature}}$ was the learning rate for feature update and was set to 0.01 except for the simulations shown in Figure 3G, in which $\alpha_{\mathrm{feature}}$ decreased over time according to 0.01/(1 + time-step/1000).

Agent was initially placed at the relief state. The IR-system-specific values for all the actions were initialized to 0, and the weight vector for the approximate SR system-specific value function $\mathbf{w}$ was initialized to $\mathbf{0}$. The SR matrix $\mathbf{M}$ was initialized to the identity matrix, i.e., the features corresponding to the own actions were initialized to 1 and all the other features were initialized to 0. The value of the time discount factor ($\gamma = 0.5$) was the same as the one used in the previous work (Sakai et al., 2022), and the ranges of the learning rates $\alpha_{\mathrm{IR}+}$, $\alpha_{\mathrm{IR}-}$, $\alpha_{\mathrm{SR}+}$, and $\alpha_{\mathrm{SR}-}$ were also determined in reference to the value of the learning rate used in that study (0.1). However, the value of the inverse temperature used in the simulations shown in the main text and figures ($\beta = 10$) was larger than the value used in the previous work ($\beta = 1$) (Sakai et al., 2022). If β was set to 1 in our model, the proportion that the agent was at the obsession state was rather high (≥ around 0.4 on average) in any combinations examined in Figure 3A-D (($\alpha_{\mathrm{SR}+}$, $\alpha_{\mathrm{SR}-}$, $\alpha_{\mathrm{IR}+}$, $\alpha_{\mathrm{IR}-}$) = (0, 0, 0.1, 0.1), (0.1, 0.1, 0, 0), (0.05, 0.05, 0.05, 0.05), and (0.09, 0.01, 0.01, 0.09)), as shown in Supplementary Figure S1.

Figure S1
Behavior of agent in the case with small degree of choice exploitation over exploration. Behavior of agent with ($\alpha_{\mathrm{SR}+}$, $\alpha_{\mathrm{SR}-}$, $\alpha_{\mathrm{IR}+}$, $\alpha_{\mathrm{IR}-}$) = (0, 0, 0.1, 0.1) (A), (0.1, 0.1, 0, 0) (B), (0.05, 0.05, 0.05, 0.05) (C), and (0.09, 0.01, 0.01, 0.09) (D) in the case where the inverse temperature parameter β, representing the degree of choice exploitation over exploration, was set to 1 (c.f., Figure 3A-D show the case where β was set to 10).

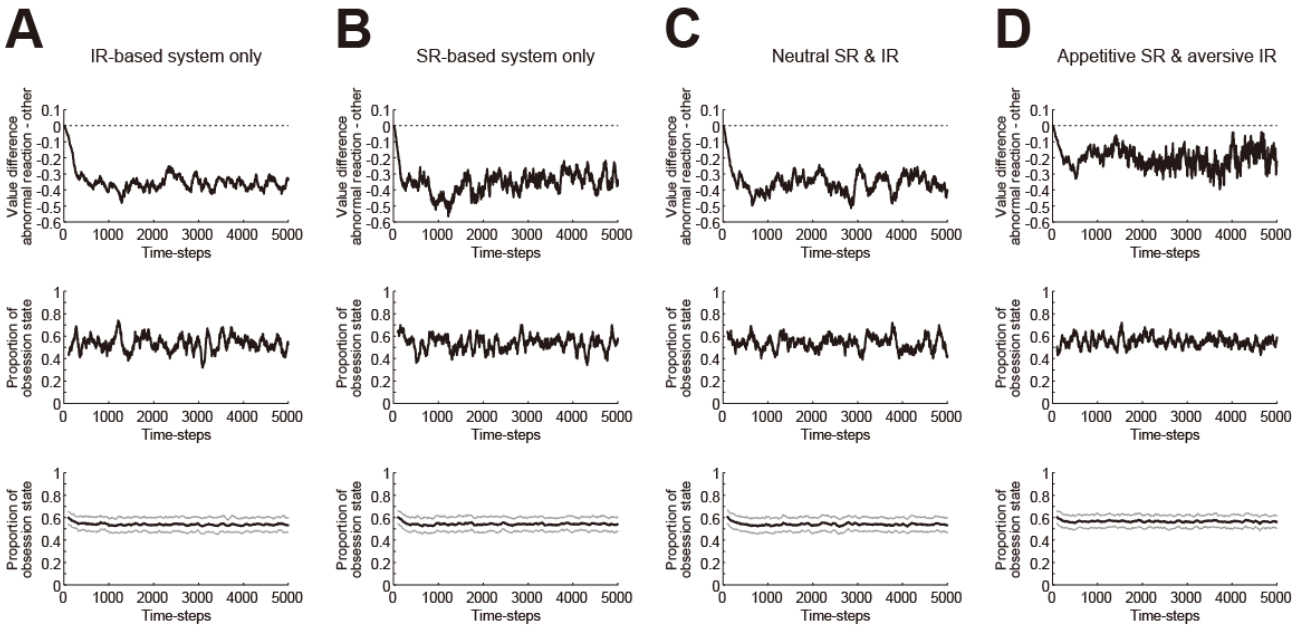

Figure S1

Simulation and fitting of the two-stage decision task

We simulated the behavior of the agent having a coupled SR- and IR-based systems in the two-stage decision task consisting of 201 trials (Daw et al., 2011). There were two choice options at the first stage, and selection of one of them led to either of two pairs of second-stage options with 70% or 30% whereas selection of the other first-stage option led to either of the two pairs with 30% and 70%. Selection of one of the second-stage options led to reward (size 1) or no-reward. The probability of reward for each second-stage option was determined according to Gaussian random walk with reflecting boundaries at 0.25 and 0.75, and it was implemented in a way that pseudo normal random numbers (mean 0, SD 0.025) was added to the probability at every trial and the probability was reflected at 0.25 and 0.75.

At every trial, choice was made in the soft-max manner with the degree of exploitation over exploration (inverse temperature) (β) set to 5 at both stages. 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 $O_k(2)$ were chosen options for the first and second stage at the k-th trial, respectively, and $V(O)$ was the value of option O (i.e., average of SR- and IR-system specific values). γ was the time discount factor, and it was set to 1 (i.e., no temporal discounting). Then, after reward (R = 1 or 0) was determined, TD-RPE for the second stage was calculated: $\delta_k(2) = R + 0 - V(O_k(2))$, where $V(O_k(2))$ reflected $\delta_k(1)$-based updates of the weights of the approximate SR system-specific value function. The IR-based system-specific value of each first- and second-stage option was initialized to 0 and

updated based on the TD-RPEs. The SR-based system had the SR of the first- and second-stage options, and the SR-based system-specific value was leaned through TD-RPE-based updates of the weights of the approximate value function. The SR was initialized to the one that assumed the random policy on second-stage choice and incorporated the actual stage-transition probabilities. At every trial, after stage transition occurred and second-stage option was determined, SR features for the chosen first-stage option was updated based on the TD errors with the learning rate 0.05. We considered agent that had appetitive SR- and aversive IR-based systems and agent that effectively had SR-based system only, for which the learning rates for both systems from positive and negative TD-RPEs were set to (α_{SR+}, α_{SR-}, α_{IR+}, α_{IR-}) = (0.27, 0.03, 0.03, 0.27) and (0.3, 0.3, 0, 0), respectively, common to both states, and conducted 100 simulations for each agent.

Generated choice sequences were individually fitted by the RL model weighing model-based and model-free controls developed in the original study (Daw et al., 2011). The RL model had seven free parameters: a_1 and a_2 (learning rates for the first and second stages), b_1 and b_2 (inverse temperatures for the first and second stages), λ (eligibility trace, representing the degree with which the model-free value of first-stage option was updated based on the TD-RPE at the second stage), ρ (perseveration or switching at first-stage choices), and w (degree (weight) of model-based control, with the degree (weight) of model-free control set to $(1-w)$). Since a_1, a_2, λ, and w should normally be between 0 and 1, these variables were transformed into $\mathrm{atanh}(2x-1)$, where x represents a_1, a_2, λ, or w, in the fitting process so that the range of exploration was effectively bound to [0 1]. For each choice sequence, we explored a set of parameters that maximized the log likelihood (Daw, 2011) by using the "fminsearch" function of MATLAB, with the initial parameter values set as (a_1, a_2, b_1, b_2, λ, ρ) = (0.3, 0.3, 5, 5, 0.2, 0.2) and w = 0.2, 0.5, and 0.8 (i.e., these three cases were examined) and the 'MaxFunEvals' and 'MaxIter' options of this function set to both 10000. Among the three cases with different initial values of w, we selected a case that gave the maximum log likelihood, and if the "fminsearch" function terminated with "exitflag" 1 (meaning normal termination) in that case, the corresponding data (value of w) was included in the following analysis. This was true for 97 out of 100 simulations for both types of agent. We compared the obtained distribution of w for each type of agent by using the Wilcoxon rank-sum test (Mann Whitney U test).

We also simulated the behavior of the original RL model (Daw et al., 2011), as well as the behavior of a modified model with imbalanced eligibility trace, in the two-stage decision task. As for the original model, the parameters were set to (a_1, a_2, b_1, b_2, λ, ρ, w) = (0.3, 0.3, 5, 5, 0.4, 0.2, 0.5). The modified model had two parameters, λ_1 and λ_2, representing the eligibility trace for positive (non-negative) and negative TD-RPEs, respectively, instead of the single parameter λ in the original model, and they were set to (λ_1, λ_2) = (0.4, 0) to represent imbalanced trace; the other parameters were set to the same values as those set for the original model. 100 simulations were conducted for each model, and generated choice sequences were individually fitted by the original model in the same manner as in the case of the agent with coupled SR- and IR-based systems described above. For each choice

sequence, among the three cases of fitting with different initial values of w, we again selected a case that gave the maximum log likelihood, and this time the "fminsearch" function terminated with "exitflag" 1 in that case for all of the 100 simulations for both models. We again compared the obtained distribution of w for each model by using the Wilcoxon rank-sum test.

Precisely speaking, our implementation of the RL model developed in (Daw et al., 2011) and its variant, used for the fitting and/or simulations mentioned above, could be slightly different from the original one. Specifically, we could not find how the model-based system estimated the transition probabilities when the frequencies of past occurrence of both types of transitions exactly matched in the original study, and we assumed that the model-based system estimated them as 50% and 50% in such situations.

Software and statistics

Simulations were conducted by using MATLAB, and its functions "rand" and "randn" were used to implement probabilities and pseudorandom numbers. Standard deviation (SD) of simulation data was normalized by n. The Wilcoxon rank-sum test (Mann Whitney U test) was conducted by using R (function "wilcox.test").

Code availability

Codes to generate the data presented in the figures will be made available at: https://github.com/kenjimoritagithub/sr101

References

Daw ND (2011) Trial-by-trial data analysis using computational models. In: Decision Making, Affect, and Learning, Attention and Performance XXIII (Delgado M, Phelps EA, Robbins TW, eds): Oxford University Press.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204-1215.

Dayan P (1993) Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Computation 5:613-624.

Dolan RJ, Dayan P (2013) Goals and habits in the brain. Neuron 80:312-325.

Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306:1940-1943.

Garvert MM, Dolan RJ, Behrens TE (2017) A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. Elife 6:e17086.

Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND (2016) Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife 5:https://doi.org/10.7554/eLife.11305.

Groman SM, Keistler C, Keip AJ, Hammarlund E, DiLeone RJ, Pittenger C, Lee D, Taylor JR (2019) Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. Neuron 103:734-746.e733.

Hikida T, Kimura K, Wada N, Funabiki K, Nakanishi S (2010) Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. Neuron 66:896-907.

Iino Y, Sawada T, Yamaguchi K, Tajiri M, Ishii S, Kasai H, Yagishita S (2020) Dopamine D2 receptors in discrimination learning and spine enlargement. Nature 579:555-560.

Kravitz AV, Tye LD, Kreitzer AC (2012) Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat Neurosci 15:816-818.

Lee SJ, Lodder B, Chen Y, Patriarchi T, Tian L, Sabatini BL (2021) Cell-type-specific asynchronous modulation of PKA by dopamine in learning. Nature 590:451-456.

Lei W, Jiao Y, Del Mar N, Reiner A (2004) Evidence for differential cortical input to direct pathway versus indirect pathway striatal projection neurons in rats. J Neurosci 24:8289-8299.

Lu J, Cheng Y, Xie X, Woodson K, Bonifacio J, Disney E, Barbee B, Wang X, Zaidi M, Wang J (2021) Whole-Brain Mapping of Direct Inputs to Dopamine D1 and D2 Receptor-Expressing Medium Spiny Neurons in the Posterior Dorsomedial Striatum. eNeuro 8:ENEURO.0348-0320.2020.

Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ (2017) The successor representation in human reinforcement learning. Nat Hum Behav 1:680-692.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936-1947.

Morita K, Shimomura K, Kawaguchi Y (2022) Opponent learning with different representations in the cortico-basal ganglia circuits. bioRxiv https://doi.org/10.1101/2021.10.29.466375.

Ohmura Y, Iwami K, Chowdhury S, Sasamori H, Sugiura C, Bouchekioua Y, Nishitani N, Yamanaka A, Yoshioka M (2021) Disruption of model-based decision making by silencing of serotonin neurons in the dorsal raphe nucleus. Curr Biol 31:2446-2454.e2445.

Robbins TW, Vaghi MM, Banca P (2019) Obsessive-Compulsive Disorder: Puzzles and Prospects. Neuron 102:27-47.

Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND (2017) Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol 13:e1005768.

Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND (2021) Neural evidence for the successor representation in choice evaluation. bioRxiv https://doi.org/10.1101/2021.08.29.458114.

Sakai Y, Sakai Y, Abe Y, Narumoto J, Tanaka SC (2022) Memory trace imbalance in reinforcement and punishment systems can reinforce implicit choices leading to obsessive-compulsive behavior. Cell Rep 40:111275.

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593-1599.

Shimomura K, Kato A, Morita K (2021) Rigid reduced successor representation as a potential mechanism for addiction. Eur J Neurosci 53:3768-3790.

Skvortsova V, Palminteri S, Pessiglione M (2014) Learning to minimize efforts versus maximizing rewards: computational principles and neural correlates. J Neurosci 34:15621-15630.

Stachenfeld KL, Botvinick MM, Gershman SJ (2017) The hippocampus as a predictive map. Nat Neurosci 20:1643-1653.

Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction (Second Edition). Cambridge, MA: MIT Press.

Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L (2012) Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci 15:1281-1289.

Voon V, Baek K, Enander J, Worbe Y, Morris LS, Harrison NA, Robbins TW, Rück C, Daw N (2015) Motivation and value influences in the relative balance of goal-directed and habitual behaviours in obsessive-compulsive disorder. Transl Psychiatry 5:e670.

Wall NR, De La Parra M, Callaway EM, Kreitzer AC (2013) Differential innervation of direct- and indirect-pathway striatal projection neurons. Neuron 79:347-360.

Worbe Y, Palminteri S, Savulich G, Daw ND, Fernandez-Egea E, Robbins TW, Voon V (2016) Valence-dependent influence of serotonin depletion on model-based choice strategy. Mol Psychiatry 21:624-629.