Title

Difficulty in Cessation of Undesired Habits: Goal-Based Reduced Successor Representation and Reward Prediction Errors

Short title

Mechanisms for the Difficulty in Cessation of Undesired Habits

Authors

Kanji Shimomura[1+], Ayaka Kato[2,3,4+], & Kenji Morita[1,5*]

[+] These authors contribute equally to this work.

Affiliations

[1] Physical and Health Education, Graduate School of Education, The University of Tokyo

[2] Department of Life Sciences, Graduate School of Arts and Sciences, The University of Tokyo

[3] Laboratory for Circuit Mechanisms of Sensory Perception, RIKEN Center for Brain Science

[4] Research Fellowship for Young Scientists, Japan Society for the Promotion of Science

[5] International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo

**Corresponding author*

Kenji Morita, Ph.D.

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

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

E-mail: morita@p.u-tokyo.ac.jp

Author contributions

Conceptualization: K.M., K.S., A.K.; Formal Analysis: K.S.; Supervision: K.M.; Validation: A.K.; Visualization: K.S.; Writing – Original Draft Preparation: K.M.; Writing – Review & Editing: K.M., A.K., K.S..

Funding

K.M. is 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/). A.K. is supported by Grant-in-Aid for JSPS Fellows (No. 19J12156) of the Japan Society for the Promotion of Science (https://www.jsps.go.jp/english/).

Acknowledgements

The authors thank Mr. Hirokazu Hatta for literature search on computational models of addiction.

Conflicts of Interest

A.K. is an employee of CureApp, Inc, Japan.

Abstract

Difficulty in cessation of drinking, smoking, or gambling, even with strong intention, has been widely recognized. Reasons for this, and whether there are reasons common to substance and non-substance reward, remain elusive. We present a computational model of common potential mechanisms underlying the difficulty in resisting habitual behavior to obtain reward. Consider that a person has long been regularly taking a series of actions leading to a purchase of alcohol, cigarette, or betting ticket without any hesitation. Referring to the recently suggested representation of states by their successors in human reinforcement learning as well as the dimension reduction in state representations in the brain, we assumed that the person has acquired a rigid representation of states along the series of habitual actions by the discounted future occupancy of the final successor state, namely, the rewarded goal state, under the established non-resistant policy. Then, we show that if the person takes a different policy to resist temptation of habitual behavior, negative reward prediction error (RPE) is generated when s/he makes "No-Go" decisions whereas no RPE occurs upon "Go" decisions, and a large positive RPE is generated upon eventually reaching the goal, given that the state representation acquired under the non-resistant policy is so rigid that it does not easily change. In the cases where the states are instead represented in the punctate manner or by the discounted future occupancies of all the states (i.e., by the genuine successor representation), negative and positive RPEs are generated upon "No-Go" and "Go" decisions, respectively, whereas no or little RPE occurs at the goal. We suggest that these RPEs, especially the large positive RPE generated upon goal reaching in the case with the goal-based reduced successor representation, might underlie the difficulty in cessation of undesired habitual or addictive behavior to obtain substance and non-substance reward.

Author Summary

Many people try to stop drinking, smoking, or gambling, but fail it. Why? In case of drinking or smoking, alcohol or nicotine could invade the brain and affect the neural circuits. But such substance-based explanations obviously do not hold for gambling or video gaming. A conceivable explanation, common for substance and non-substance, is that such behavior has become a habit, which is so rigidly established that it could not be changed. However, it has been shown that even those who are suffering from severe drug addiction can behave in a goal-directed, rather than habitual, manner in the sense that they can exhibit intact sensitivity to changes in the value of action outcomes. Meanwhile, recent work suggests that humans may develop a subtler "habit", where a particular type of internal representation of states (situations), rather than action itself, becomes rigidly formed. Here we show, through computational modeling, that if a similar type of, but dimension-reduced, state representation is formed, when people try to resist long-standing reward-obtaining behavior but eventually fail and reach the rewarded goal, a large positive "prediction error" of rewards would arise, and discuss that it might underlie the difficulty in cessation of undesired habitual behavior.

Introduction

Difficulty in cessation of drinking, smoking, or gambling, even with strong intention, has been widely recognized. Reasons for this, and whether there are reasons common to substance and non-substance reward, remain elusive. Although much effort has been devoted to developing clinical programs including technology-based therapies (e.g. [1, 2]; reviewed in [3, 4]), the lack of mechanistic understanding of the undesired habit is an obstacle for further improvement. Computational modeling has become a powerful approach to elucidating the mechanisms of psychiatric disorders including addiction [5-8]. However, it appears that relatively less focus has been given to non-substance, compared to substance, addiction, although there have been proposals (e.g., [9-11]). Also, while previous computational studies appear to typically focus on the difficulty in withdrawal from severe addiction, including the problems of relapse, presumably a larger population has established milder, stable additive behavior for years but then decides to quit such behavior because it potentially causes health or socioeconomic problems rather than because the behavior itself is immediately problematic as pathological addiction. In the present study, we explored possible computational mechanisms for the difficulty in resisting such stably established habitual behavior to obtain reward, with the following four streams of findings and suggestions in mind:

(1) *Involvement of the dopamine (DA) system in both substance and non-substance addiction*

The DA system has been suggested to be crucially involved in substance addiction [12], possibly through drug-induced DA acting as a fictitious RPE that cannot be canceled out by predictions [13, 14]. However, there have also been implications of possible involvements of the DA system in non-substance addiction, such as possible relations of medicines of Parkinson disease to pathological gambling [15, 16].

(2) *Goal-directed and habitual behavior and their neural substrates, and their relations to addiction*

It has been suggested that there are two behavioral processes, namely, goal-directed and habitual behavior, which are sensitive or insensitive to changes in outcome values and/or action-outcome contingencies, respectively [17-19]. They are suggested to be hosted by distinct corticostriatal circuits, specifically, those including ventral/dorsomedial striatum (or caudate) and those including dorsolateral striatum (or putamen), respectively [20-22], where ventral-to-dorsal spiral influences have been anatomically suggested [23, 24]. Computationally, goal-directed and habitual behavior have been suggested to correspond to model-based reinforcement learning (RL) and model-free RL, respectively ([25]; but see [26] for a critique of model-free RL as a model of habitual behavior). It has been suggested that addiction can be caused by impaired goal-directed and/or excessive habitual control [27, 28]. This is supported by multitudes of animal experiments, and there also exist findings in humans in line with this [29]. However, it has also been shown that human addicts often do show goal-directed behavior, such as those sensitive to outcome devaluation [30]. Also, there have been proposals of many different possible causes for addiction, including those related to each of the two control systems and/or in their interactions, the way of state representation, or the hierarchical organization of the learning system (e.g., [9, 11, 31-33]).

(3) *Intermediate of goal-directed and habitual behavior through successor representation of states*

A great mystery had been that how model-based and model-free RLs, whose typical algorithms are so different in formulae, can be both hosted by corticostriatal-DA circuits, different parts of which should still share basic architectures. Recent work [34, 35] has provided a brilliant potential solution to this by proposing that certain types of goal-directed (model-based) behavior, having sensitivity to changes in outcome values, can be achieved through a particular type of state representation called the successor representation [36], combined with the ever-suggested representation of RPE by DA [37, 38]. In the successor representation, individual states are represented by a sort of closeness to their successor states, or more accurately, by time-discounted cumulative future occupancies of these states. Behavior based on this representation is not fully goal-directed, having difficulty in revaluation of state transition or policy, which has been demonstrated in actual human behavior [39] referred to as "subtler, more cognitive notion of habit" by the authors [39]. Successor representation and value update based on it have been suggested to be implemented in the prefrontal/hippocampus-dorsomedial/ventral striatum circuits [34, 40, 41], while circuits including dorsolateral striatum might implement habitual or model-free behavior through "punctate" representation of states.

(4) *Sustained DA response to predictable reward, possibly related to state representation*

The original experiments that led to the proposal of representation of RPE by DA [37, 38] have shown that DA response to reward disappears after monkeys repeatedly experienced the stimulus(-action)-

reward association and the reward presumably became predictable for them. However, sustained, and often ramping, dopamine signals to/towards (apparently) predictable reward has been widely observed in recent years [42-49]. There are a number of possible accounts for such sustained DA signals, positing that they represent RPE [47, 50-54] or something different from RPE [42, 43, 45, 46, 48, 49] or both [44, 55]. Of particular interest to our present work, one hypothesis [50] suggests that sustained (ramping) DA signals might represent sustained RPE generated due to imperfect approximation of value function in the system using representation of states by low-dimensional features.

Referring to these different streams of findings and suggestions, we propose a computational model of potential mechanisms underlying the difficulty in resisting undesired habitual behavior to obtain reward. We have found that a dimension-reduced successor representation of states leads to the generation of a distinct pattern of RPE, which might particularly underlie such difficulty.

Results

Goal-based reduced successor representation of states under non-resistant policy

We modeled a person's series of actions to obtain a particular reward, such as alcohol, nicotine, or non-substance such as betting ticket or social interaction, by a series of modeled person's actions on a sequence of states from the start state to the goal state, where the reward is given (Fig. 1A). At each state except for the goal state, the person can take either of two actions, "Go": proceed to the next state, and "No-Go": stay at the same state (as considered in our previous work [52] in a different context). We considered a case that the person has long been regularly taking behavior to obtain the reward without resisting temptation. In the model, it corresponds to that the person has long experienced transitions towards the rewarded goal according to a policy that takes only "Go" at any state, which we refer to as the Non-Resistant policy. We assumed that, through such long-standing experiences of behavior according to the Non-Resistant policy, the person has established a particular state representation, where each state is represented by the discounted future occupancy of the final successor state, namely, the rewarded goal state, under that policy. Specifically, we considered a single (i.e., scalar) feature x and assumed that the k-th state, S_k (k = 1, ..., n; S_1 is the start state and S_n is the goal state), is represented by:

$$x(S_k) = \gamma^{n-k}, \qquad \text{(Eq. 1)}$$

where γ is the time discount factor. The number of states (n) was set to 10, and the time discount factor

(γ) was assumed to be 0.97, resulting in that the discounted value at the start state was $0.97^9 \approx 0.76$ times of the value at the goal, unless otherwise mentioned (Fig. 1B) (see the Methods for rationale for these parameter values).

This representation, which we will refer to as the goal-based representation, can be said to be a dimension-reduced version of successor representation; in the genuine successor representation [34-36], every state is represented by a vector of expected cumulative discounted future state occupancies for all the states, whereas in the above goal-based representation, every state is represented by the discounted future occupancy of only the goal state. Because the genuine successor representation requires the number of features equal to the number of states, dimension reduction has been considered (c.f., [56-58]). Given the general suggestion of dimension reduction in state representations in the brain [59, 60], it would be conceivable that the brain adopts dimension-reduced versions of successor representation, such as the goal-based representation assumed above. Notably, the state value function under the Non-Resistant policy in the case of the assumed structure of state transitions and rewards (Fig. 1A) can be precisely represented as a linear function of the scalar feature of the assumed goal-based representation. Specifically, the state value for S_k is given by

$$V_{\text{Non-Resistant}}(S_k) = R_n\gamma^{n-k} = R_n x(S_k), \qquad \text{(Eq. 2)}$$

where R_n is the reward value obtained at the goal state, which was assumed to be 1. Therefore, the assumed goal-based representation can be said to be a minimal representation for achieving accurate state values, and it would thus be conceivable that such a representation has been acquired through long-lasting behavior.

Regarding implementation in the brain, a finding that the BOLD signal in the ventromedial prefrontal cortex and hippocampus was negatively correlated with the distance to the goal in a navigation task [61] appears to be in line with such a goal-based representation; if those regions engaged predominantly in the genuine successor representation in that task, their overall activity may not show a monotonic increase towards the goal. It is conceivable that the genuine successor representation can be encoded in the hippocampus [40], but the reduced goal-based representation can become dominant through intensive training on a particular task or through long-lasting habitual behavior towards a particular goal. Another study [62] has shown that the BOLD signal in the posterior hippocampus was positively correlated with the path distance to the goal (increased as the path became farther) during travel periods whereas it was negatively correlated with an interaction between the distance and direction to the goal (increased as the path became closer and more direct) at decision points (and prior studies potentially in line with either of these results are cited therein [63-66]). The goal-based representation that we assumed can potentially be in line with the activity at decision points, rather than during travel periods, in that study.

RPEs under resistant policy, with state representation under non-resistant policy

We then modeled a situation where the person decides to attempt cessation of the habitual reward-obtaining behavior by assuming that the person starts to take a new policy, referred to as the Resistant policy, in which not only "Go" but also "No-Go" action is chosen with a certain probability, $P_{\text{No-Go}}$, at each state preceding the goal. Crucially, we assumed that the goal-based state representation has been established so rigidly through long-standing behavior under the Non-Resistant policy that the representation does not change after the person changes the policy to the Resistant policy. We therefore assumed that the person tries to approximate the state value function under the new, Resistant policy by a linear function of the abovementioned scalar feature, $x(S_k)$, with a coefficient w:

$$V_{\text{Resistant}}(S_k) \approx V^{approximate}{}_{\text{Resistant}}(S_k) = wx(S_k), \qquad \text{(Eq. 3)}$$

by updating the coefficient w using the temporal-difference (TD) RPE at every time step:

$$\delta = R(S(t)) + \gamma wx(S(t+1)) - wx(S(t)), \qquad \text{(Eq. 4)}$$

where $S(t)$ and $S(t+1)$ are the states at time t and $t+1$, respectively, and if $S(t)$ is the goal state, the term $\gamma wx(S(t+1))$ is dropped. $R(S(t))$ is the reward value obtained at $S(t)$, which was assumed to be 0 except for the goal state. Specifically, w was assumed to be updated as follows:

$$w \rightarrow w + \alpha x(S(t))\delta, \qquad \text{(Eq. 5)}$$

where α is the learning rate, which was set to 0.5. This way of linear function approximation and TD-RPE-based update [67, 68] has been typically assumed in neuro-computational models and is considered to be implementable through synaptic plasticity depending on DA, which represents δ, and presynaptic activity, which represents $x(S(t))$ [34, 37]. The initial value of w was set to R_n (= 1), with which the approximate value function exactly matches the true value function under the Non-Resistant policy; notably, we did not simulate the person's behavior under the Non-Resistant policy, but we instead just assumed this initial value of w and simulated the person's behavior under the Resistant policy only. The probability of "No-Go" choice ($P_{\text{No-Go}}$) was set to 0.75; later we also describe results with different values of $P_{\text{No-Go}}$.

We then examined RPEs generated upon each decision, "Go" or "No-Go", at each state before the goal state or upon reaching the goal state. Figure 2Aa shows a single simulation example of RPEs generated at each state in the first episode. In this episode, the person chose "No-Go" once at S_3, twice at S_5 and S_6, four times at S_4, seven times at S_1, S_2, and S_9, nine times at S_8, and never at S_7. The blue crosses indicate RPEs generated upon "Go" decisions, whereas the red crosses indicate the means of RPEs generated upon "No-Go" decisions, and the black cross indicates RPE generated at the goal state. The magenta circles indicate the summation of RPEs generated upon "No-Go" decisions at the same states. As shown in the figure, when the person chose "No-Go", negative RPEs were generated, whereas theoretically no RPE is generated upon choosing "Go" (though tiny numerical errors existed (the same applies throughout)), and when the person eventually reached the rewarded goal state, a positive RPE was generated. Figure 2Ab shows the mean and standard deviation across simulations. The same features as observed in the example simulation are observed.

Figure 2B shows the over-episode change of the coefficient w of the approximate value function at the end of each episode, averaged across simulations. As shown in the figure, w decreases from its initial value, R_n (= 1), and becomes (almost) stationary, meaning that the negative and positive

RPE-based updates become overall balanced. We examined RPEs after the coefficient w becomes nearly stationary, in particular, in the 25th episode. Figure 2C shows the results averaged across simulations. As shown in the figure, negative RPEs were generated upon "No-Go" decisions, whereas theoretically no RPE is generated upon "Go" decisions, and a large positive RPE was generated upon reaching the rewarded goal state. We also examined how the amplitudes of RPEs change over episodes, specifically for RPEs generated upon "Go" or "No-Go" decisions at the start state and RPE generated at the goal state (Fig. 2D). As shown in the figure, after a few initial episodes, the amplitudes, averaged across simulations, become nearly stationary.

We also examined the cases with different parameters, in particular, with the probability of "No-Go" choice ($P_{\text{No-Go}}$) varied over 0.5, 0.75 (assumed above), and 0.9, and the time discount factor (γ) varied over 0.95, 0.97 (assumed above), and 0.99. Figure 3 shows the results for RPEs in the 25th episode (Fig. 3A) and over-episode changes of RPEs upon "Go" or "No-Go" decisions at the start state and RPE at the goal state (Fig. 3B) (the center panels of Fig. 3A,B show the same data as shown in Fig. 2C,D, respectively). As shown in the figures, basic features mentioned above are largely preserved over these parameter ranges. The sustained negative RPEs generated upon "No-Go" decisions are considered to potentially deteriorate the propensity to resist temptation of habitual behavior to obtain reward. Also, the sustained large positive RPE generated at the rewarded goal state is considered to potentially reinforce a final reward-taking behavior, although it is not included in our model. Therefore, we propose that these RPEs can be potential mechanisms underlying the difficulty in resisting habitual behavior to obtain reward.

If the person continues to take the Resistant policy for a number of times, it can be expected that the goal-based representation of states established under the Non-Resistant policy slowly changes, and gradually approaches the representation under the Resistant policy, particularly through TD learning of state representation itself [58, 69]. We thus examined how RPEs become if it occurs. Specifically, we came back to the original parameters considered in Fig. 2, i.e., $P_{\text{No-Go}} = 0.75$ and $\gamma = 0.97$, and conducted simulations in the same way as above, except that this time we also updated the scalar feature of the state (i.e., $x(S(t))$) at every time step by using the TD error of the goal-based representation:

$$\delta_{\text{GR}} = 0 + \gamma x(S(t+1)) - x(S(t)). \qquad \text{(Eq. 6)}$$

Specifically, the scalar feature was updated as follows:

$$x(S(t)) \longrightarrow x(S(t)) + \alpha_{\text{GR}}\delta_{\text{GR}}, \qquad \text{(Eq. 7)}$$

where α_{GR} is the learning rate for this update and was set to 0.05, except for the goal state, for which the TD error of the goal-based representation should be theoretically 0 and thus no update was implemented. Figure 4A shows the scalar feature of each state (i.e., $x(S_k)$) after 50, 100, and 200 episodes (black dotted, dashed, and solid lines, respectively), averaged across simulations, in comparison to the original ones (gray line). As shown in the figure, the curve became steeper as episodes proceeded. This is considered to reflect that longer time is required, on average, for goal reaching under the Resistant policy than under the Non-Resistant policy and thus the expected discounted future occupancy of the goal state should be smaller for the Resistant policy. Figure 4B shows the RPEs generated in the 200th episode, averaged across simulations, and Figure 4C shows the over-episode changes in the RPEs generated upon "Go" or "No-Go" decisions at the start state and the RPE generated at the goal state. As shown in these figures, the large positive RPE generated upon goal reaching observed in the case with the original state representation (Fig. 2C) gradually decreased, while positive RPEs with smaller amplitudes gradually appeared upon "Go" decisions in the states other than the goal. This indicates that if the large positive RPE upon goal reaching is especially harmful, it can be resolved if the person does not give up resisting temptation, even with the help of clinical intervention such as alternative reward upon "No-Go" choices, until the state representation considerably changes.

Comparison to the cases with punctate representation or genuine successor representation of states

For comparison, we considered a case where each state is represented in the "punctate" manner. For this case, we assumed that each state has its own state value, $V_{\text{punctate}}(S_k)$, and it is updated using TD-RPE in the punctate system, δ_{punctate}:

$$\delta_{\text{punctate}} = R(S(t)) + \gamma V_{\text{punctate}}(S(t{+}1)) - V_{\text{punctate}}(S(t)), \qquad \text{(Eq. 8)}$$

where if $S(t)$ is the goal state, the term $\gamma V_{\text{punctate}}(S(t{+}1))$ is dropped. Specifically, V_{punctate} was assumed to be updated as follows:

$$V_{\text{punctate}}(S(t)) \rightarrow V_{\text{punctate}}(S(t)) + \alpha_{\text{punctate}}\delta_{\text{punctate}}, \qquad \text{(Eq. 9)}$$

where α_{punctate} is the learning rate for the punctate system, which was set to 0.5. The initial values for the punctate state values were assumed to be:

$$V_{\text{punctate}}(S_k) = R_n \gamma^{n-k}, \qquad \text{(Eq. 10)}$$

which are the state values under the Non-Resistant policy as considered in Eq. 2.

Figure 5 shows the RPEs generated in the punctate system in the same various conditions as examined for the system with the goal-based representation (Fig. 3). Comparing these two figures, prominent differences are that whereas theoretically no RPE occurs upon "Go" decisions and a large positive RPE is generated upon goal reaching in the case with the goal-based representation, (relatively small) positive RPEs are generated upon "Go" decision and theoretically no RPE occurs at the goal in the case with punctate representation. These differences are considered to reflect different characteristics of updates done with the different ways of state representation. Specifically, in the case with the goal-based representation, only the coefficient of approximate value function was updated and the state representation established under the Non-Resistant policy was (assumed to be) unchanged, resulting in sustained mismatch between the true and approximate value functions. In contrast, in the case with punctate state representation, the value of each state was directly updated so that there is no such sustained mismatch.

We also considered a case where the states are represented by the genuine successor representation. Specifically, we assumed that each state S_k is represented by n features $x_j(S_k)$ (j = 1, ..., n) indicating the time-discounted future occupancy of S_j under the Non-Resistant policy:

$$x_j(S_k) = \gamma^{j-k}\ (j \geq k) \text{ or } 0\ (j < k),$$

and the value function under the Resistant policy is approximated by a linear function of them:

$$V_{\text{Resistant}}(S_k) \approx \Sigma_{j=1:n}\ \{w_j x_j(S_k)\}.$$

The coefficients w_j ($j = 1, ..., n$) are updated by using the TD RPE:

$$\delta = R(S(t)) + \gamma\Sigma_{j=1:n}\ \{w_j x_j(S(t+1))\} - \Sigma_{j=1:n}\ \{w_j x_j(S(t))\},$$

where the middle term including $S(t+1)$ is dropped if $S(t)$ is the goal state, according to the following rule:

$$w_j \rightarrow w_j + \alpha x_j(S(t))\delta.$$

The initial values of w_j were set to 0 for $j = 1, ..., n-1$ and R_n (= 1) for $j = n$, with which the approximate value function exactly matches the true value function under the Non-Resistant policy. Figure 6A shows the RPEs generated at each state in the 25th episode, and Figure 6B shows the over-episode changes in the RPEs generated upon "Go" or "No-Go" decisions at the start state and the RPE generated at the goal state, both with the original parameters considered in Fig. 2, i.e., $P_{\text{No-Go}} = 0.75$ and $\gamma = 0.97$. As shown in the figures, the patterns of RPEs are similar to those in the case of punctate representation (the center panels of Fig. 5A,B) and differ from those in the case of the goal-based representation. Figure 6C shows the coefficients w_j of the approximate value function after the 1st episode (Fig. 6Ca) and 25th episode (Fig. 6Cb), and Figure 6D shows the over-episode changes of the coefficients for the features corresponding to the start state (red line), the state preceding the goal (S_9) (blue line), and the goal state (black line). As shown in these figures, the coefficients for the features corresponding to the states preceding the goal became negative. It is considered that because of these negative coefficients, the true value function under the Resistant policy could be well approximated even by a linear function of the features (discounted occupancies) under the Non-Resistant policy.

Discussion

Previous studies have suggested that the DA system is involved in both substance and non-substance addictions [12-16]. Also, while impaired goal-directed and/or excessive habitual control have been suggested for addiction [27-29], human addicts often show intact sensitivity to outcome devaluation [30], and many different possible causes for addiction have also been proposed (e.g., [9, 11, 31-33]). Meanwhile, recent neuroscience research has suggested that partially goal-directed but partially habitual behavior is realized through successor representation coupled with DA-RPE [34, 35, 39], and also that sustained DA response to predictable reward might occur depending on state representation [50]. Referring to these different streams of suggestions, we have proposed a computational model of potential mechanisms underlying the difficulty in resisting habitual behavior to obtain reward. In particular, in the model consisting of a series of state transitions towards the rewarded goal, we have shown that negative RPEs upon "No-Go" decisions and a large positive RPE upon goal reaching are generated in the system with the goal-based representation established under the Non-Resistant policy, whereas negative and positive RPEs upon "No-Go" and "Go" decisions, respectively, but no RPE at the goal, are generated in the system with punctate representation or genuine successor representation. Below we discuss how these RPEs, especially the large positive RPE upon goal reaching in the case with the goal-based representation, could underlie the difficulty in cessation of long-standing habitual behavior to obtain reward. Successor representation is a neurally implementable way of partially model-based RL, but one of its critical drawbacks is policy-dependence [34, 39, 70]. Dimension reduction in state representation in the brain is generally suggested [59, 60], but it is inevitably accompanied by the risk of inaccuracy. The present work proposes that these negative sides are related to the difficulty in the cessation of undesired habits.

Possible effects of generated RPEs on behavior

The negative RPEs upon "No-Go" decisions appeared with all the examined types of state representations, as well as positive RPEs upon "Go" decisions appeared in the punctate or genuine successor representation system, are considered to potentially deteriorate the propensity to resist temptation of habitual behavior to obtain reward. For example, if the probabilities of "Go" and "No-Go" choices are not fixed as assumed in the present work but determined by action preferences, which are updated by RPEs according to the actor-critic method [71-73], it is expected that the preference of "No-Go" action will decrease (and the preference of "Go" action will increase in the punctate or genuine successor representation case) so that the probability of "No-Go" choice will decrease, returning back to the original Non-Resistant policy. Also, the negative RPEs upon "No-Go" decisions might potentially cause subjective negative feelings in real humans, given the suggestion that subjective momentary happiness of humans could be explained by reward expectations and RPEs [74].

On the other hand, the sustained large positive RPE generated at the rewarded goal state in the system with the goal-based representation might potentially cause subjective positive feelings. Also, whether linked to subjective feelings or not, the large positive RPE upon goal reaching is considered to potentially reinforce reward-taking behavior, although it is not included in our model. In our model, as in the previous model considering successor representation [34], RPE generated at state $S(t)$ (Eq. 4) was assumed to contain the reward value obtained at that state ($R(S(t))$), rather than at the state after transition ($R(S(t+1))$), and also the time-discounted estimated value of the state after transition ($\gamma wx(S(t+1))$). If RPE-based update of action preference is introduced into the model, it is natural to consider updating, using the RPE at $S(t)$, the preference of action at $S(t)$ that causes the transition to $S(t+1)$. But then, the preference of "Go" action at the state preceding the goal state will be updated by RPE at that state, which is theoretically 0, rather than by the positive RPE at the goal state. However, if we assume multi-step eligibility trace (c.f., [67, 68]) for actions, i.e., assume that RPE is used to update the preference of not only the immediate action but also, to a lesser degree, the preceding action, the preference of "Go" action at the state preceding the goal state is expected to be increased by the positive RPE at the goal state.

We also think of another possible way for the large positive RPE at the rewarded goal state to affect behavior. As mentioned in the Results, RPEs generated in the case with the goal-based representation continuously update the coefficient w of the approximate value function, but negative and positive updates are overall balanced so that w can remain stationary across episodes. However, as mentioned in the Introduction, it is suggested that there exist multiple value learning systems in the brain, with the system employing successor representation residing in the prefrontal/hippocampus-dorsomedial/ventral striatum circuits, whereas another system adopting punctate representation might locate in the circuits including dorsolateral striatum. Moreover, there are anatomical suggestions of ventral-to-dorsal spiral influences in the striatum-midbrain system [23, 24], and a theoretical proposal that the effect of drug-induced DA accumulates through the spiraling circuit and causes undesired compulsive drug taking in long-term addicts [32]. Given these, and if the prefrontal/hippocampus-dorsomedial/ventral striatum circuits encode the goal-based representation, rather than the genuine successor representation, RPEs generated in that system might not only train the values in itself but also affect the system with punctate representation. If this is the case, the large positive RPEs generated at the goal state in the goal-based representation system is expected to increase the value of the goal state, or of a reward-taking action, in the punctate system, resulting in that resisting temptation could ironically cause more compulsive reward taking. In the worst case, such large positive RPEs, coming from the outside of the punctate system, could potentially even act as fictitious RPEs that cannot be canceled out by predictions within the punctate system and thereby causes unbounded value increase and compulsion, similarly to what has been suggested for drug-induced DA [13]. But this last case is not very likely, given that the suggested ventral-to-dorsal spiral influences [23, 24] in fact indicate projections of more ventral parts of striatum to more dorsal parts of midbrain rather than projections of more ventral parts of midbrain to more dorsal parts of striatum, the latter of which would cause direct DA inflow. Nonetheless, we speculate that direct DA invasion from the goal-based representation system to the punctate system might still potentially occur if there exist dorsolateral striatum-projecting DA neurons that receive (direct or indirect) inputs only from the ventral/dorsomedial striatum (so that cancelation by inputs from the dorsolateral striatum cannot occur) and the amplitude of the positive RPE is so large.

Possible experimental validation, and clinical implication

Our model predicts that distinct patterns of RPEs are generated at each state leading to the rewarded goal in the systems with the goal-based representation (Fig. 3) and punctate representation (Fig. 5) or genuine successor representation (Fig. 6), which are presumably encoded by DA released in different parts of the striatum and cortex. This prediction can potentially be tested by fMRI experiments and model-based analyses [75, 76]. There are, however, two anticipated problems. First, we assumed that the goal-based representation under the Non-Resistant policy has been so rigidly established through long-standing, e.g., years of, habitual behavior that it does not change after the policy is changed to the Resistant policy. It would be not easy to create a situation that can mimic such a long-standing habitual behavior in laboratory experiments, even with some (e.g., weeks of) pre-training before entering the scanner. Second, we aimed to model a series of actions leading to reward such as alcohol, nicotine, or betting ticket, which typically take dozens to tens of minutes and are presumably accompanied with some time discounting (see the Methods for rationale for the parameter values assumed for this aim). In laboratory experiments consisting of trials lasting for seconds with monetary rewards, time discounting may have less effects. However, this could be overcome, given that time discounting was actually measured in experimental task in humans [77].

From clinical perspectives, it is essential to know whether the phenomena described by the present model, either those with the goal-based representation or those with punctate or genuine successor representation, or both, actually occur in people who are trying to resist long-standing behavior to obtain reward, and whether the generated RPEs indeed underlie, at least partly, the difficulty in cessation of such behavior. A possible way is to conduct brain imaging for those people executing a task that simulates their daily struggles against reward-obtaining behavior, including failures to resist temptation, although it is again necessary to overcome the second problem mentioned above. If it is then suggested that the large positive RPE upon goal reaching generated in the system with the goal-based representation is an important cause of the difficulty, a possible intervention that is potentially effective is to provide alternative reward (physical, social, or internal) upon "No-Go" decisions until the state representation changes and approaches the one under the Resistant policy.

Methods

Equations and parameters used are described in the Results. As described there, we set the number of states (n) to 10, and the time discount factor (γ) was varied over 0.97 ± 0.02, resulting in that the value at the start state was 0.95^9 (≈ 0.63), 0.97^9 (≈ 0.76), or 0.99^9 (≈ 0.91) times of the value at the goal. We assumed 10 states because it seems intuitively reasonable to assume that the long-standing daily behavior to obtain a particular reward, such as going to a favorite pub for a beer after work, consists of around several to 10 distinct actions, e.g., clean the desktop, wear the jacket, wait for and get on the elevator, walk to the subway station, wait for and get on a train, walk to the pub, call the waitstaff, and order the beer. These series of actions would typically take dozens to tens of minutes. Given this, we determined the abovementioned range of time discount factor in reference to a study [78], which examined temporal discounting for video gaming and found that the subjective value of video gaming 1 hour later was on average around 0.65 ~ 0.8 times of the value of immediate video gaming. Notably, however, the temporal discounting reported in that study appears to have near flat tails, indicating that it would not be well approximated by exponential functions, whereas we assumed exponential discounting.

In order to examine average behavior of the model across simulations, simulations were conducted 100 times for each condition. Among the 100 simulations, there were likely to be simulations, where "No-Go" choice was not taken at some state(s) at some episode(s). Such simulations, different from case to case, were not included in the calculations of the average and standard deviation of RPEs across simulations. There were also likely to be simulations, where "No-Go" choice was taken more than once at some state(s) at some episode(s). In such cases, generated RPEs were first averaged within an episode, and that value (i.e., a single value for each simulation) was used for the calculations of the average and standard deviation of RPEs across simulations. Simulations and figure drawing were conducted by using Python (3.7.2) and R (4.0.0), respectively. Program codes for generating all the data presented in the figures are planned to be made available at the GitHub after acceptance for journal publication.

Acknowledgements

The authors thank Mr. Hirokazu Hatta for literature search on computational models of addiction.

Figure legends

Figure 1

Schematic diagram of the model and the assumed goal-based representation of states under the Non-Resistant policy. **(A)** Schematic diagram of the model, adapted, with alterations, from Fig. 1 of [52]. **(B)** Assumed representation of states by the discounted future occupancy of the final successor state, namely, the rewarded goal state, under the Non-Resistant policy, in which only "Go" is chosen at any state, except for the goal state.

Figure 2

RPEs generated under the Resistant policy, with the goal-based representation under the Non-Resistant policy. **(A)** **(a)** A single simulation example of RPEs generated at each state in the first episode. The blue crosses indicate RPEs generated upon "Go" decisions, whereas the red crosses indicate the means of RPEs generated upon "No-Go" decisions, and the black cross indicates RPE generated at the goal state. The magenta circles indicate the summation of RPEs generated upon "No-Go" decisions at the same states. **(b)** The average across simulations. The error bars indicate ± standard deviation (SD); this is also applied to the following figures. **(B)** Over-episode change of the coefficient *w* of the approximate value function at the end of each episode, averaged across simulations, and the shading indicates ± SD (this is also applied to the following figures); the assumed initial value (*w* = 1) is also plotted at episode = 0 with SD = 0. **(C)** RPEs generated at each state in the 25th episode, where the coefficient *w* has become nearly stationary, averaged across simulations (± SD). **(D)** The changes of RPEs over episodes, averaged across simulations (± SD). RPEs generated upon "Go" decisions (blue) and "No-Go" decisions (mean (red) and summation (magenta) per episode) at the start state, and RPE generated at the goal state (black).

Figure 3

Cases with different parameters. The probability of "No-Go" choice ($P_{\text{No-Go}}$) was set to 0.5, 0.75 (the value assumed in Fig. 2), and 0.9. The time discount factor (γ) was set to 0.95, 0.97 (the value assumed in Fig. 2), and 0.99. **(A)** RPEs generated at each state in the 25th episode, averaged across simulations (± SD). Notations are the same as those in Fig. 2C (blue: "Go", red cross: "No-Go" mean, magenta circle: "No-Go" sum, black: goal), and the center panel shows the same data as shown in Fig. 2C. **(B)** Over-episode changes of RPEs upon "Go" and "No-Go" decisions at the start state, and RPE generated at the goal state, averaged across simulations (± SD). Notations are the same as those in Fig. 2D (blue: "Go" at the start, red: "No-Go" mean at the start, magenta: "No-Go" sum at the start, black: at the goal), and the center panel shows the same data as shown in Fig. 2D.

Figure 4

RPEs generated under the Resistant policy, in the case where the goal-based state representation itself slowly changed and approached the goal-based representation under the Resistant policy. The probability of "No-Go" choice ($P_{\text{No-Go}}$) was set to 0.75, and the time discount factor (γ) was set to 0.97. **(A)** Scalar feature of each state (i.e., $x(S_k)$) after 50, 100, and 200 episodes (black dotted, dashed, and solid lines, respectively), averaged across simulations (± SD), in comparison to the original ones (gray line) that are the same as those shown in Fig. 1B. **(B)** RPEs generated at each state in the 200th episode, averaged across simulations (± SD). Notations are the same as those in Fig. 2C (blue: "Go", red cross: "No-Go" mean, magenta circle: "No-Go" sum, black: goal). **(C)** Over-episode changes of RPEs upon "Go" and "No-Go" decisions at the start state, and RPE generated at the goal state, averaged across simulations (± SD). Notations are the same as those in Fig. 2D (blue: "Go" at the start, red: "No-Go" mean at the start, magenta: "No-Go" sum at the start, black: at the goal).

Figure 5

RPEs generated in the system with punctate representation of states. The conditions, parameters, and notations are the same as those in Fig. 3. **(A)** RPEs generated at each state in the 25th episode, averaged across simulations (± SD) (blue: "Go", red cross: "No-Go" mean, magenta circle: "No-Go" sum, black: goal). **(B)** Over-episode changes of RPEs upon "Go" and "No-Go" decisions at the start state, and RPE generated at the goal state, averaged across simulations (± SD) (blue: "Go" at the start, red: "No-Go" mean at the start, magenta: "No-Go" sum at the start, black: at the goal).

Figure 6

Case with genuine successor representation. The probability of "No-Go" choice ($P_{\text{No-Go}}$) was set to 0.75, and the time discount factor (γ) was set to 0.97. **(A)** RPEs generated at each state in the 25th episode, averaged across simulations (± SD). Notations are the same as those in Fig. 2C (blue: "Go", red cross: "No-Go" mean, magenta circle: "No-Go" sum, black: goal). **(B)** Over-episode changes of RPEs upon "Go" and "No-Go" decisions at the start state, and RPE generated at the goal state, averaged across simulations (± SD). Notations are the same as those in Fig. 2D (blue: "Go" at the start, red: "No-Go" mean at the start, magenta: "No-Go" sum at the start, black: at the goal). **(C)** Coefficients w_j of the approximate value function after the 1st episode (a) and 25th episode (b), averaged across simulations (± SD). **(D)** Over-episode changes of the coefficients w_j for the features corresponding to the start state (red line), the state preceding the goal (S_9) (blue line), and the goal state (black line), averaged across simulations (± SD).

References

1. Gustafson DH, McTavish FM, Chih MY, Atwood AK, Johnson RA, Boyle MG, et al. A smartphone application to support recovery from alcoholism: a randomized clinical trial. JAMA Psychiatry. 2014;71(5):566-72. doi: 10.1001/jamapsychiatry.2013.4642. PubMed PMID: 24671165; PubMed Central PMCID: PMCPMC4016167.

2. Kato A, Tanigawa T, Satake K, Nomura A. Efficacy of the Ascure Smoking Cessation Program: Retrospective Study. JMIR Mhealth Uhealth. 2020;8(5):e17270. Epub 2020/05/14. doi: 10.2196/17270. PubMed PMID: 32406856.

3. Newman MG, Szkodny LE, Llera SJ, Przeworski A. A review of technology-assisted self-help and minimal contact therapies for drug and alcohol abuse and smoking addiction: is human contact necessary for therapeutic efficacy? Clin Psychol Rev. 2011;31(1):178-86. Epub 2010/10/21. doi: 10.1016/j.cpr.2010.10.002. PubMed PMID: 21095051.

4. Haskins BL, Lesperance D, Gibbons P, Boudreaux ED. A systematic review of smartphone applications for smoking cessation. Transl Behav Med. 2017;7(2):292-9. doi: 10.1007/s13142-017-0492-2. PubMed PMID: 28527027; PubMed Central PMCID: PMCPMC5526818.

5. Montague PR, Dolan RJ, Friston KJ, Dayan P. Computational psychiatry. Trends Cogn Sci. 2012;16(1):72-80. doi: S1364-6613(11)00251-8 [pii] 10.1016/j.tics.2011.11.018. PubMed PMID: 22177032.

6. Wang XJ, Krystal JH. Computational psychiatry. Neuron. 2014;84(3):638-54. Epub 2014/11/05. doi: 10.1016/j.neuron.2014.10.018. PubMed PMID: 25442941; PubMed Central PMCID: PMCPMC4255477.
7. Huys QJ, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci. 2016;19(3):404-13. doi: 10.1038/nn.4238. PubMed PMID: 26906507; PubMed Central PMCID: PMCPMC5443409.
8. Kato A, Kunisato Y, Katahira K, Okimura T, Yamashita Y. Computational Psychiatry Research Map (CPSYMAP): a New Database for Visualizing Research Papers. bioRxiv. 2020;https://doi.org/10.1101/2020.06.30.181198. doi: https://doi.org/10.1101/2020.06.30.181198.
9. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol Rev. 2007;114(3):784-805. doi: 10.1037/0033-295X.114.3.784. PubMed PMID: 17638506.
10. Piray P, Keramati MM, Dezfouli A, Lucas C, Mokri A. Individual differences in nucleus accumbens dopamine receptors predict development of addiction-like behavior: a computational approach. Neural Comput. 2010;22(9):2334-68. doi: 10.1162/NECO_a_00009. PubMed PMID: 20569176.
11. Ognibene D, Fiore VG, Gu X. Addiction beyond pharmacological effects: The role of environment complexity and bounded rationality. Neural Netw. 2019;116:269-78. Epub 2019/05/08. doi:

10.1016/j.neunet.2019.04.022. PubMed PMID: 31125913; PubMed Central PMCID: PMCPMC6581592.

12. Berke JD, Hyman SE. Addiction, dopamine, and the molecular mechanisms of memory. Neuron. 2000;25(3):515-32. doi: 10.1016/s0896-6273(00)81056-9. PubMed PMID: 10774721.
13. Redish AD. Addiction as a computational process gone awry. Science. 2004;306(5703):1944-7. doi: 10.1126/science.1102384. PubMed PMID: 15591205.
14. Keiflin R, Janak PH. Dopamine Prediction Errors in Reward Learning and Addiction: From Theory to Neural Circuitry. Neuron. 2015;88(2):247-63. doi: 10.1016/j.neuron.2015.08.037. PubMed PMID: 26494275; PubMed Central PMCID: PMCPMC4760620.
15. Dodd ML, Klos KJ, Bower JH, Geda YE, Josephs KA, Ahlskog JE. Pathological gambling caused by drugs used to treat Parkinson disease. Arch Neurol. 2005;62(9):1377-81. Epub 2005/07/11. doi: 10.1001/archneur.62.9.noc50009. PubMed PMID: 16009751.
16. Voon V, Hassan K, Zurowski M, Duff-Canning S, de Souza M, Fox S, et al. Prospective prevalence of pathologic gambling and medication association in Parkinson disease. Neurology. 2006;66(11):1750-2. doi: 10.1212/01.wnl.0000218206.20920.4d. PubMed PMID: 16769956.
17. Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998;37(4-5):407-19. doi: 10.1016/s0028-3908(98)00033-1. PubMed PMID: 9704982.
18. 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.

19. 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.

20. Corbit LH, Muir JL, Balleine BW. The role of the nucleus accumbens in instrumental conditioning: Evidence of a functional dissociation between accumbens core and shell. J Neurosci. 2001;21(9):3251-60. PubMed PMID: 11312310; PubMed Central PMCID: PMCPMC6762583.

21. Yin HH, Knowlton BJ, Balleine BW. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci. 2004;19(1):181-9. doi: 10.1111/j.1460-9568.2004.03095.x. PubMed PMID: 14750976.

22. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci. 2005;22(2):513-23. doi: 10.1111/j.1460-9568.2005.04218.x. PubMed PMID: 16045504.

23. Haber SN, Fudge JL, McFarland NR. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J Neurosci. 2000;20(6):2369-82. PubMed PMID: 10704511; PubMed Central PMCID: PMCPMC6772499.

24. Joel D, Weiner I. The connections of the dopaminergic system with the striatum in rats and

primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience. 2000;96(3):451-74. doi: S0306-4522(99)00575-8 [pii]. PubMed PMID: 10717427.

25. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8(12):1704-11. doi: nn1560 [pii] 10.1038/nn1560. PubMed PMID: 16286932.

26. Dezfouli A, Balleine BW. Habits, action sequences and reinforcement learning. Eur J Neurosci. 2012;35(7):1036-51. doi: 10.1111/j.1460-9568.2012.08050.x. PubMed PMID: 22487034; PubMed Central PMCID: PMCPMC3325518.

27. Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci. 2005;8(11):1481-9. doi: 10.1038/nn1579. PubMed PMID: 16251991.

28. Everitt BJ, Robbins TW. Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. Annu Rev Psychol. 2016;67:23-50. Epub 2015/08/07. doi: 10.1146/annurev-psych-122414-033457. PubMed PMID: 26253543.

29. Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife. 2016;5. Epub 2016/03/01. doi: 10.7554/eLife.11305. PubMed PMID: 26928075; PubMed Central PMCID: PMCPMC4786435.

30. Hogarth L, Lam-Cassettari C, Pacitti H, Currah T, Mahlberg J, Hartley L, et al. Intact goal-directed control in treatment-seeking drug users indexed by outcome-devaluation and Pavlovian to

instrumental transfer: critique of habit theory. Eur J Neurosci. 2019;50(3):2513-25. Epub 2018/07/25. doi: 10.1111/ejn.13961. PubMed PMID: 29787620.

31. Redish AD, Jensen S, Johnson A. A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci. 2008;31(4):415-37; discussion 37-87. doi: 10.1017/s0140525x0800472x. PubMed PMID: 18662461.

32. Keramati M, Gutkin B. Imbalanced decision hierarchy in addicts emerging from drug-hijacked dopamine spiraling circuit. PLoS One. 2013;8(4):e61489. Epub 2013/04/24. doi: 10.1371/journal.pone.0061489. PubMed PMID: 23637842; PubMed Central PMCID: PMCPMC3634778.

33. Keramati M, Durand A, Girardeau P, Gutkin B, Ahmed SH. Cocaine addiction as a homeostatic reinforcement learning disorder. Psychol Rev. 2017;124(2):130-53. Epub 2017/01/16. doi: 10.1037/rev0000046. PubMed PMID: 28095003.

34. 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.

35. Gershman SJ. The Successor Representation: Its Computational Logic and Neural Substrates. J Neurosci. 2018;38(33):7193-200. Epub 2018/07/13. doi: 10.1523/JNEUROSCI.0151-18.2018. PubMed PMID: 30006364; PubMed Central PMCID: PMCPMC6096039.

36. Dayan P. Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Computation. 1993;5(4):613-24.

37. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 1996;16(5):1936-47. PubMed PMID: 8774460.

38. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275(5306):1593-9. PubMed PMID: 9054347.

39. 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.

40. 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.

41. Garvert MM, Dolan RJ, Behrens TE. A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. Elife. 2017;6. Epub 2017/04/27. doi: 10.7554/eLife.17086. PubMed PMID: 28448253; PubMed Central PMCID: PMCPMC5407855.

42. Howe MW, Tierney PL, Sandberg SG, Phillips PE, Graybiel AM. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature. 2013;500(7464):575-9. doi: 10.1038/nature12475. PubMed PMID: 23913271.

43. Hamid AA, Pettibone JR, Mabrouk OS, Hetrick VL, Schmidt R, Vander Weele CM, et al. Mesolimbic dopamine signals the value of work. Nat Neurosci. 2016;19(1):117-26. doi: 10.1038/nn.4173. PubMed PMID: 26595651; PubMed Central PMCID: PMCPMC4696912.

44. Collins AL, Greenfield VY, Bye JK, Linker KE, Wang AS, Wassum KM. Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. Sci Rep. 2016;6:20231. doi: 10.1038/srep20231. PubMed PMID: 26869075.

45. Mohebi A, Pettibone JR, Hamid AA, Wong JT, Vinson LT, Patriarchi T, et al. Dissociable dopamine dynamics for learning and motivation. Nature. 2019;570(7759):65-70. Epub 2019/05/22. doi: 10.1038/s41586-019-1235-y. PubMed PMID: 31118513; PubMed Central PMCID: PMCPMC6555489.

46. Hamid AA, Frank MJ, Moore CI. Dopamine waves as a mechanism for spatiotemporal credit assignment. bioRxiv. 2019;https://doi.org/10.1101/729640. doi: https://doi.org/10.1101/729640.

47. Kim HR, Malik AN, Mikhael JG, Bech P, Tsutsui-Kimura I, Sun F, et al. A unified framework for dopamine signals across timescales. bioRxiv. 2019. doi: https://doi.org/10.1101/803437.

48. Sarno S, Beirán M, Diaz-deLeon G, Rossi-Pool R, Romo R, Parga N. Midbrain dopamine firing activity codes reward expectation and motivation in a parametric working memory task. bioRxiv. 2020;https://doi.org/10.1101/2020.05.01.071977. doi: https://doi.org/10.1101/2020.05.01.071977.

49. Guru A, Seo C, Post RJ, Kullakanda DS, Schaffer JA, Warden MR. Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. bioRxiv.

2020;https://doi.org/10.1101/2020.05.21.108886.

50. Gershman SJ. Dopamine ramps are a consequence of reward prediction errors. Neural Comput. 2014;26(3):467-71. doi: 10.1162/NECO_a_00559. PubMed PMID: 24320851.

51. Morita K, Kato A. Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits. 2014;8:36. doi: 10.3389/fncir.2014.00036.

52. Kato A, Morita K. Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. PLoS Comput Biol. 2016;12(10):e1005145. Epub 2016/10/13. doi: 10.1371/journal.pcbi.1005145. PubMed PMID: 27736881; PubMed Central PMCID: PMCPMC5063413.

53. Mikhael JG, Kim HR, Uchida N, Gershman SJ. Ramping and State Uncertainty in the Dopamine Signal. bioRxiv. 2019. doi: https://doi.org/10.1101/805366.

54. Song MR, Lee SW. Dynamic resource allocation during reinforcement learning accounts for ramping and phasic dopamine activity. Neural Netw. 2020;126:95-107. Epub 2020/03/10. doi: 10.1016/j.neunet.2020.03.005. PubMed PMID: 32203877.

55. Lloyd K, Dayan P. Tamping Ramping: Algorithmic, Implementational, and Computational Explanations of Phasic Dopamine Signals in the Accumbens. PLoS Comput Biol. 2015;11(12):e1004622. doi: 10.1371/journal.pcbi.1004622. PubMed PMID: 26699940; PubMed Central PMCID: PMCPMC4689534.

56. Gehring CA. Approximate Linear Successor Representation. Reinforcement Learning Decision Making. The multi-disciplinary conference on Reinforcement Learning and Decision Making (RLDM). 2015:http://people.csail.mit.edu/gehring/publications/clement-gehring-rldm-2015.pdf.

57. 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.

58. Gardner MPH, Schoenbaum G, Gershman SJ. Rethinking dopamine as generalized prediction error. Proc Biol Sci. 2018;285(1891). Epub 2018/11/21. doi: 10.1098/rspb.2018.1645. PubMed PMID: 30464063; PubMed Central PMCID: PMCPMC6253385.

59. Gershman SJ, Niv Y. Learning latent structure: carving nature at its joints. Curr Opin Neurobiol. 2010;20(2):251-6. Epub 2010/03/11. doi: 10.1016/j.conb.2010.02.008. PubMed PMID: 20227271; PubMed Central PMCID: PMCPMC2862793.

60. Niv Y. Learning task-state representations. Nat Neurosci. 2019;22(10):1544-53. Epub 2019/09/24. doi: 10.1038/s41593-019-0470-8. PubMed PMID: 31551597; PubMed Central PMCID: PMCPMC7241310.

61. Balaguer J, Spiers H, Hassabis D, Summerfield C. Neural Mechanisms of Hierarchical Planning in a Virtual Subway Network. Neuron. 2016;90(4):893-903. doi: 10.1016/j.neuron.2016.03.037. PubMed PMID: 27196978; PubMed Central PMCID: PMCPMC4882377.

62. Howard LR, Javadi AH, Yu Y, Mill RD, Morrison LC, Knight R, et al. The hippocampus and entorhinal cortex encode the path and Euclidean distances to goals during navigation. Curr Biol.

2014;24(12):1331-40. Epub 2014/06/05. doi: 10.1016/j.cub.2014.05.001. PubMed PMID: 24909328; PubMed Central PMCID: PMCPMC4062938.

63. Spiers HJ, Maguire EA. A navigational guidance system in the human brain. Hippocampus. 2007;17(8):618-26. doi: 10.1002/hipo.20298. PubMed PMID: 17492693; PubMed Central PMCID: PMCPMC2570439.

64. Morgan LK, Macevoy SP, Aguirre GK, Epstein RA. Distances between real-world locations are represented in the human hippocampus. J Neurosci. 2011;31(4):1238-45. doi: 10.1523/JNEUROSCI.4667-10.2011. PubMed PMID: 21273408; PubMed Central PMCID: PMCPMC3074276.

65. Viard A, Doeller CF, Hartley T, Bird CM, Burgess N. Anterior hippocampus and goal-directed spatial decision making. J Neurosci. 2011;31(12):4613-21. doi: 10.1523/JNEUROSCI.4640-10.2011. PubMed PMID: 21430161; PubMed Central PMCID: PMCPMC6622909.

66. Sherrill KR, Erdem UM, Ross RS, Brown TI, Hasselmo ME, Stern CE. Hippocampus and retrosplenial cortex combine path integration signals for successful navigation. J Neurosci. 2013;33(49):19304-13. doi: 10.1523/JNEUROSCI.1825-13.2013. PubMed PMID: 24305826; PubMed Central PMCID: PMCPMC3850045.

67. Sutton RS. Learning to predict by the methods of temporal differences. Machine learning. 1988;3:9-44.

68. Sutton RS, Barto AG. Reinforcement Learning: An Introduction (Second Edition). Cambridge,

MA: MIT Press; 2018.

69. 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.

70. Piray P, Daw, ND. A common model explaining flexible decision making, grid fields and cognitive control. bioRxiv. 2019. doi: http://dx.doi.org/10.1101/856849.

71. Barto AG, Sutton RS, Anderson CW. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on Systems, Man, and Cybernetics. 1983;13:834-46.

72. Houk J, Adams J, Barto A. A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement. In: Houk JC, Davis JL, Beiser DG, editors. Models of Information Processing in the Basal Ganglia. Cambridge, MA: MIT Press; 1995.

73. Takahashi Y, Schoenbaum G, Niv Y. Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. Front Neurosci. 2008;2(1):86-99. Epub 2008/07/09. doi: 10.3389/neuro.01.014.2008. PubMed PMID: 18982111; PubMed Central PMCID: PMCPMC2570074.

74. Rutledge RB, Skandali N, Dayan P, Dolan RJ. A computational and neural model of momentary subjective well-being. Proc Natl Acad Sci U S A. 2014;111(33):12252-7. doi: 10.1073/pnas.1407535111. PubMed PMID: 25092308; PubMed Central PMCID:

PMCPMC4143018.

75. O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci. 2007;1104:35-53. doi: annals.1390.022 [pii] 10.1196/annals.1390.022. PubMed PMID: 17416921.

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

77. Schweighofer N, Bertin M, Shishida K, Okamoto Y, Tanaka SC, Yamawaki S, et al. Low-serotonin levels increase delayed reward discounting in humans. J Neurosci. 2008;28(17):4528-32. doi: 28/17/4528 [pii] 10.1523/JNEUROSCI.4982-07.2008. PubMed PMID: 18434531.

78. Buono FD, Sprong ME, Lloyd DP, Cutter CJ, Printz DM, Sullivan RM, et al. Delay Discounting of Video Game Players: Comparison of Time Duration Among Gamers. Cyberpsychol Behav Soc Netw. 2017;20(2):104-8. Epub 2017/01/24. doi: 10.1089/cyber.2016.0451. PubMed PMID: 28118044; PubMed Central PMCID: PMCPMC5312545.

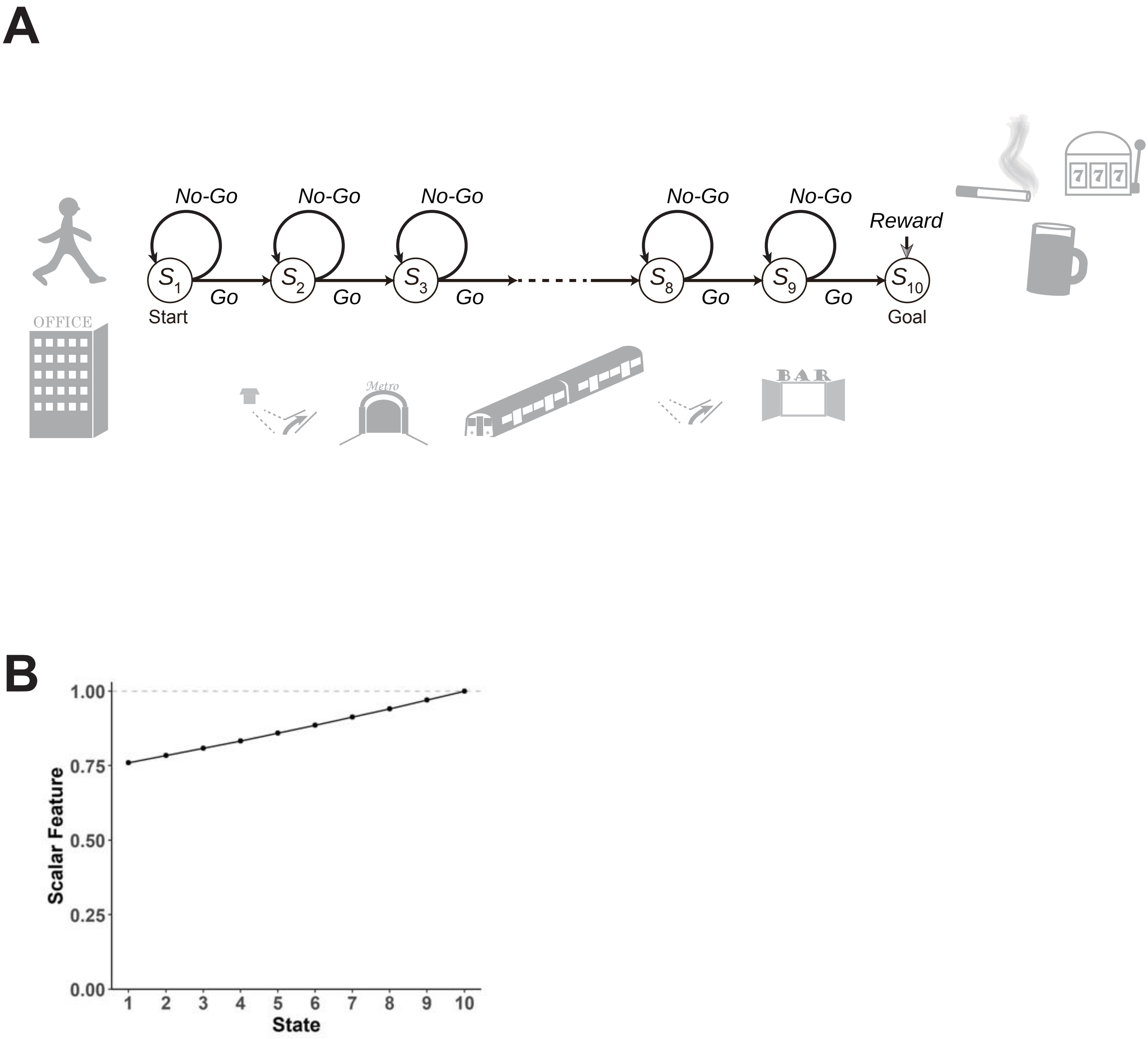
A
No-Go
No-Go
No-Go
No-Go
No-Go
Reward
S_1
S_2
S_3
S_8
S_9
S_{10}
Go
Go
Go
Go
Go
Start
Goal
OFFICE
Metro
BAR
777
B
1.00
0.75
0.50
0.25
0.00
Scalar Feature
1
2
3
4
5
6
7
8
9
10
State

Figure 1

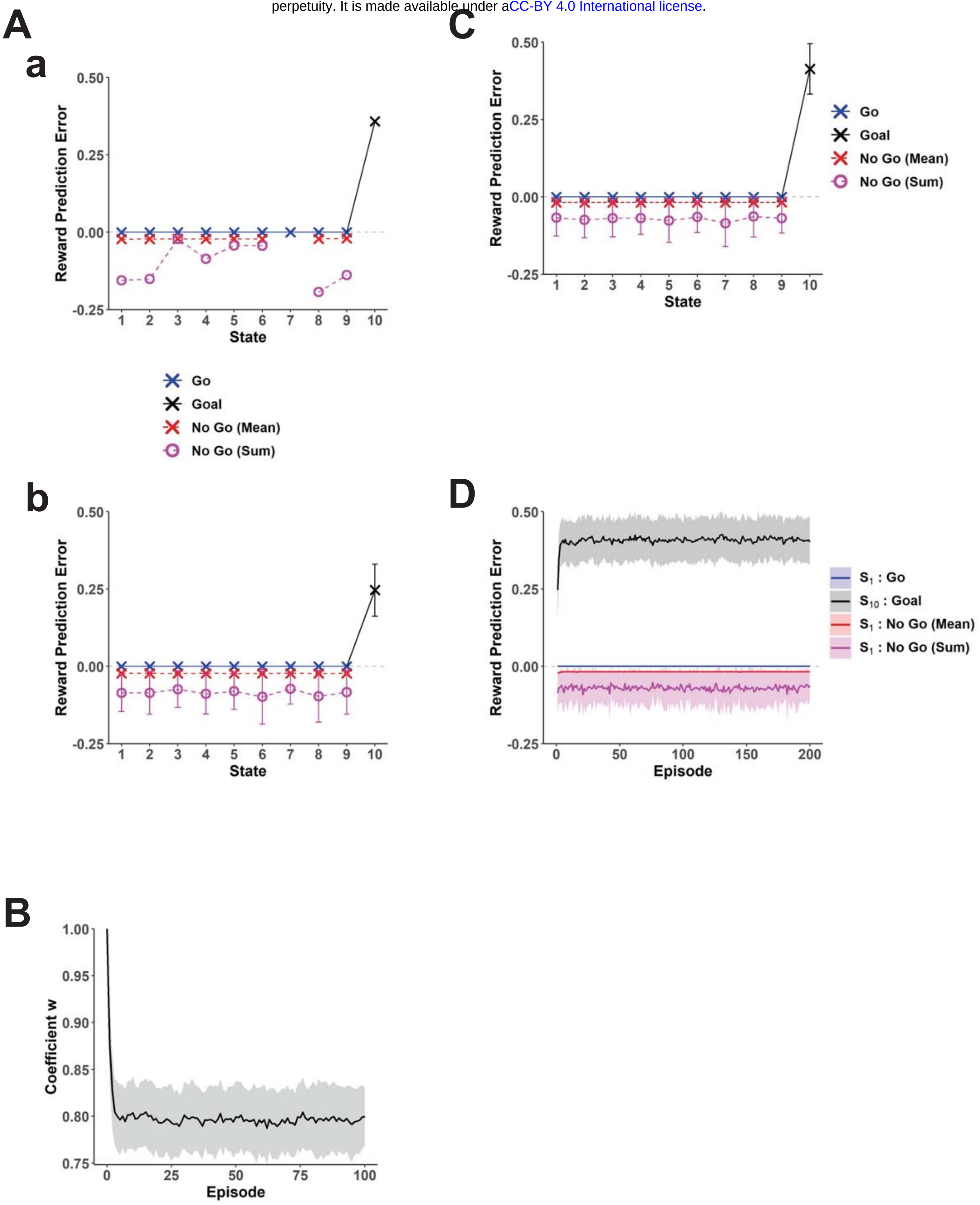
A
a
Reward Prediction Error
0.50
0.25
0.00
-0.25
1 2 3 4 5 6 7 8 9 10
State
Go
Goal
No Go (Mean)
No Go (Sum)
b
Reward Prediction Error
0.50
0.25
0.00
-0.25
1 2 3 4 5 6 7 8 9 10
State
C
Reward Prediction Error
0.50
0.25
0.00
-0.25
1 2 3 4 5 6 7 8 9 10
State
Go
Goal
No Go (Mean)
No Go (Sum)
D
Reward Prediction Error
0.50
0.25
0.00
-0.25
0 50 100 150 200
Episode
S_1 : Go
S_{10} : Goal
S_1 : No Go (Mean)
S_1 : No Go (Sum)
B
Coefficient w
1.00
0.95
0.90
0.85
0.80
0.75
0 25 50 75 100
Episode

Figure 2

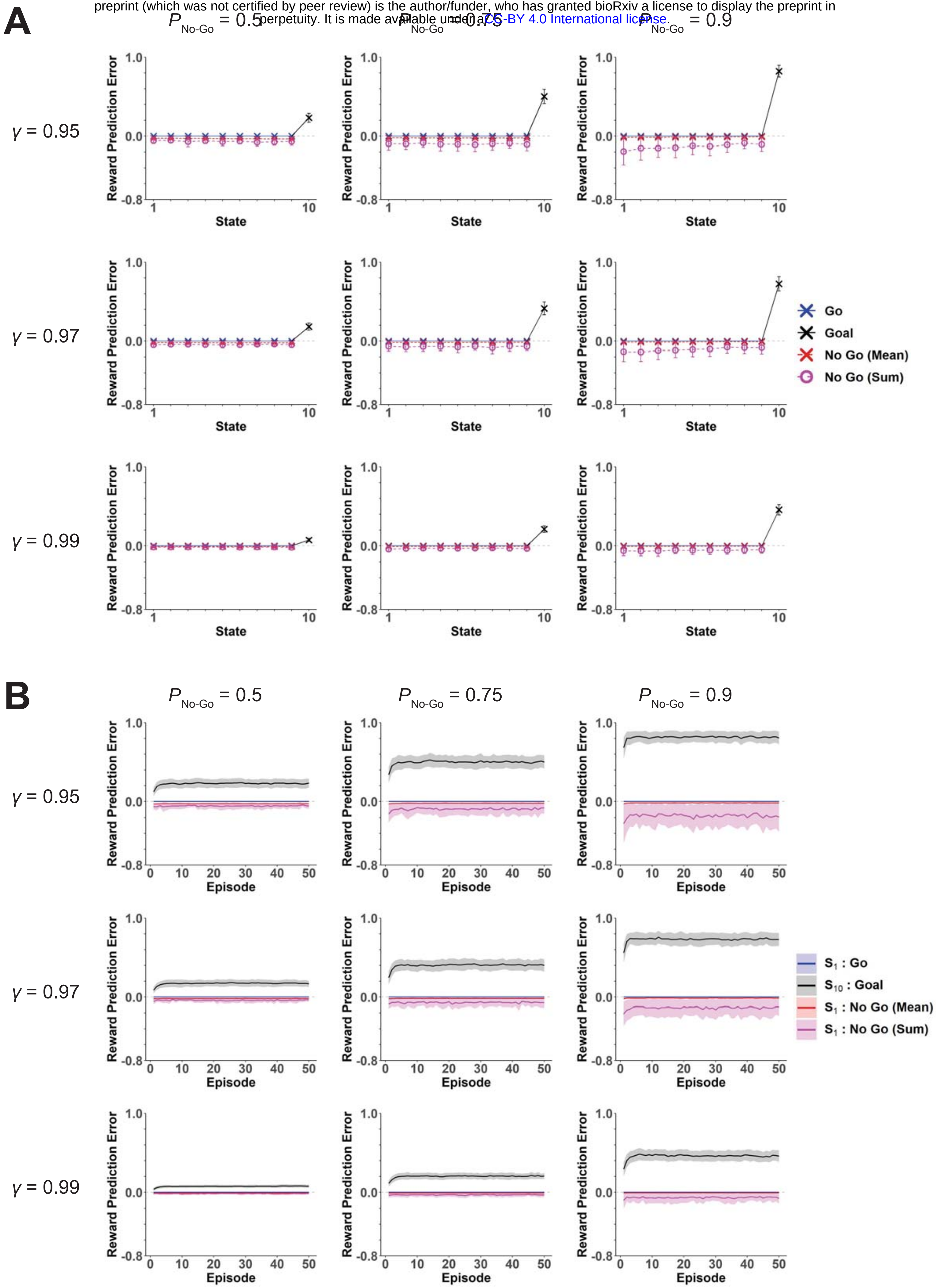
A
P No-Go = 0.5
P No-Go = 0.75
P No-Go = 0.9
γ = 0.95
γ = 0.97
γ = 0.99
Reward Prediction Error
1.0
0.0
-0.8
State
1
10
Go
Goal
No Go (Mean)
No Go (Sum)
B
P No-Go = 0.5
P No-Go = 0.75
P No-Go = 0.9
γ = 0.95
γ = 0.97
γ = 0.99
Reward Prediction Error
1.0
0.0
-0.8
Episode
0
10
20
30
40
50
S1 : Go
S10 : Goal
S1 : No Go (Mean)
S1 : No Go (Sum)

Figure 3

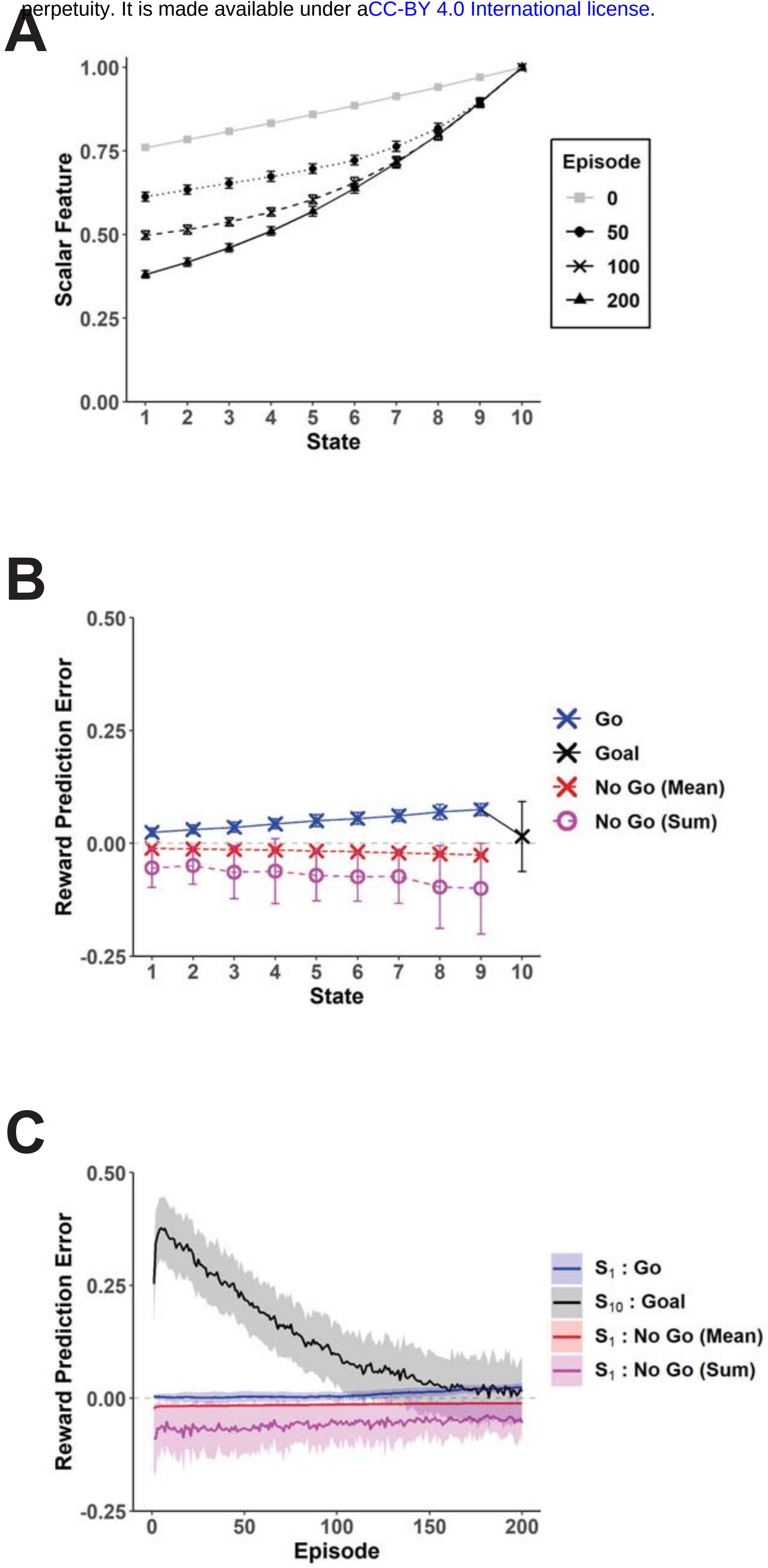
A
Scalar Feature
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
State
Episode
0
50
100
200
B
Reward Prediction Error
0.50
0.25
0.00
-0.25
1
2
3
4
5
6
7
8
9
10
State
Go
Goal
No Go (Mean)
No Go (Sum)
C
Reward Prediction Error
0.50
0.25
0.00
-0.25
0
50
100
150
200
Episode
S_1 : Go
S_{10} : Goal
S_1 : No Go (Mean)
S_1 : No Go (Sum)

Figure 4

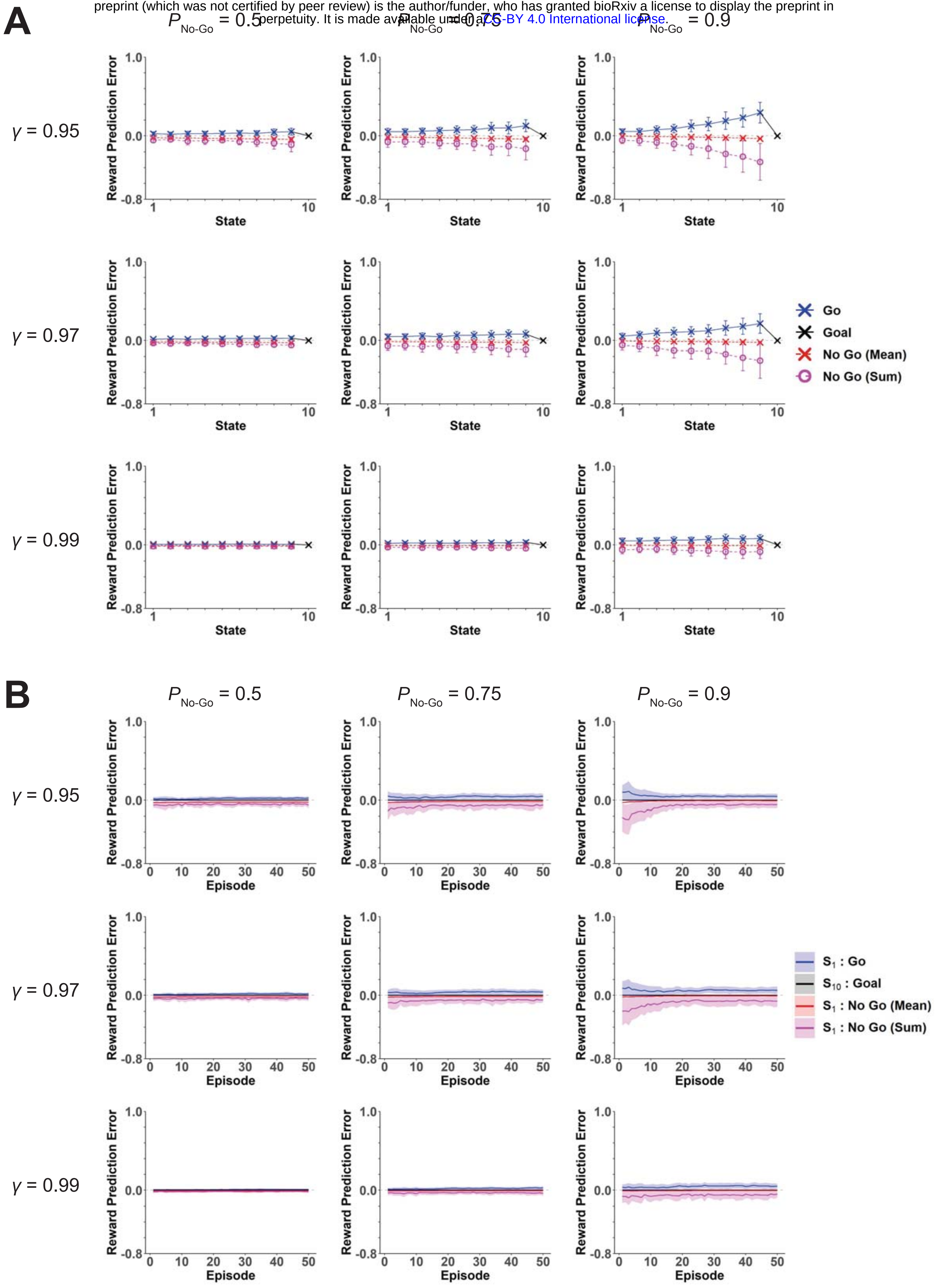
A
$P_{\text{No-Go}}$ = 0.5
$P_{\text{No-Go}}$ = 0.75
$P_{\text{No-Go}}$ = 0.9
γ = 0.95
γ = 0.97
γ = 0.99
Reward Prediction Error
1.0
0.0
-0.8
State
1
10
Go
Goal
No Go (Mean)
No Go (Sum)
B
$P_{\text{No-Go}}$ = 0.5
$P_{\text{No-Go}}$ = 0.75
$P_{\text{No-Go}}$ = 0.9
γ = 0.95
γ = 0.97
γ = 0.99
Reward Prediction Error
Episode
0
10
20
30
40
50
S_1 : Go
S_{10} : Goal
S_1 : No Go (Mean)
S_1 : No Go (Sum)

Figure 5

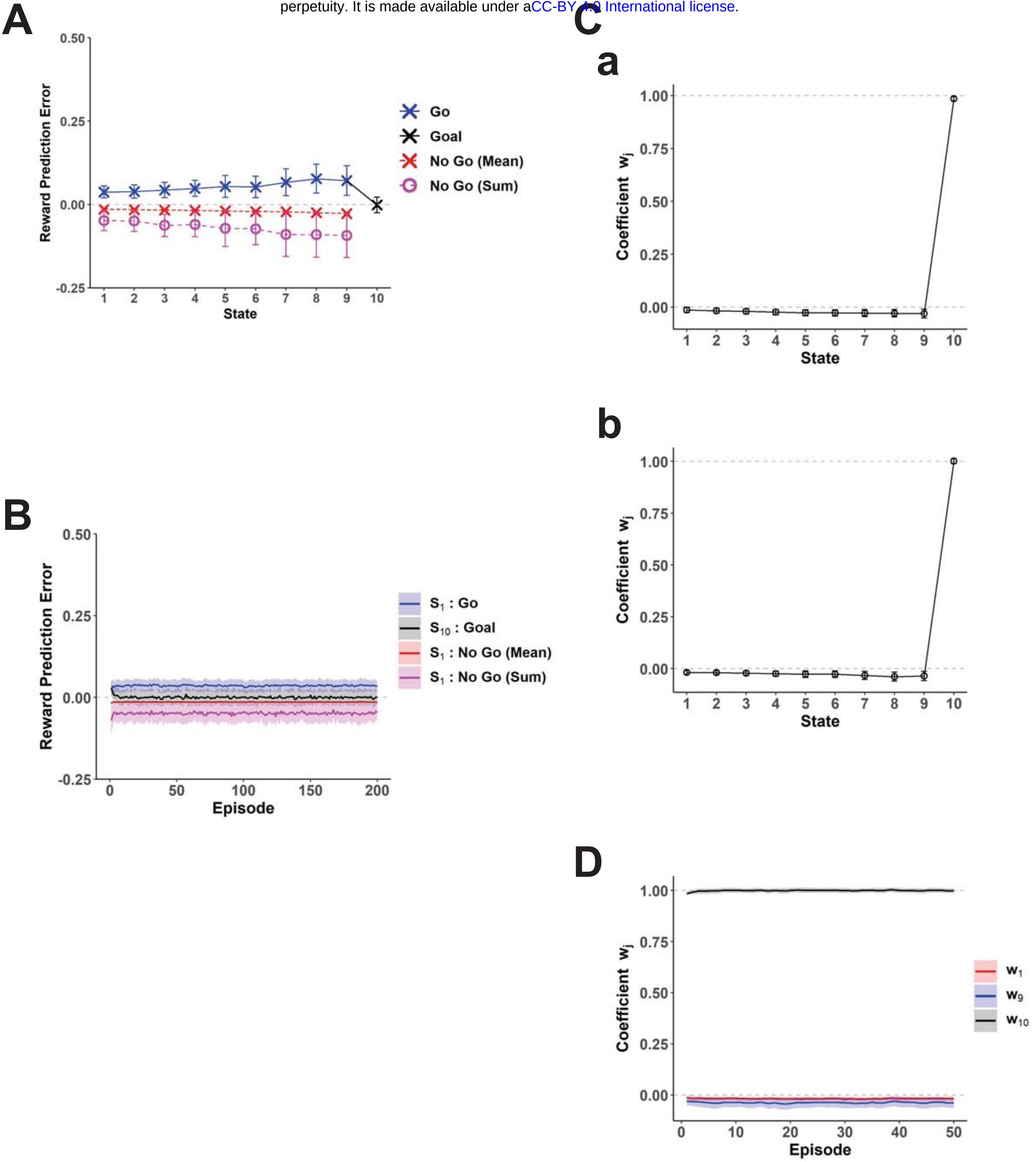
A
Reward Prediction Error
0.50
0.25
0.00
-0.25
1 2 3 4 5 6 7 8 9 10
State
Go
Goal
No Go (Mean)
No Go (Sum)
B
Reward Prediction Error
0.50
0.25
0.00
-0.25
0 50 100 150 200
Episode
S_1 : Go
S_{10} : Goal
S_1 : No Go (Mean)
S_1 : No Go (Sum)
C
a
Coefficient w_j
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10
State
b
Coefficient w_j
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10
State
D
Coefficient w_j
1.00
0.75
0.50
0.25
0.00
0 10 20 30 40 50
Episode
w_1
w_9
w_{10}

Figure 6